IDOBE: Infectious Disease Outbreak forecasting Benchmark Ecosystem

Aniruddha Adiga*
Jingyuan Chou
Anshul Chiranth
Bryan Lewis
Biocomplexity Institute, University of Virginia
Charlottesville, Virginia, USA
*aa5dw@virginia.edu

Ana I. Bento
Department of Public and Ecosystem Health Cornell University College of Veterinary Medicine
Ithaca, New York, USA

Shaun Truelove
Johns Hopkins University
Baltimore, Maryland, USA

Geoffrey Fox
Madhav Marathe
Biocomplexity Institute and Department of Computer Science, University of Virginia
Charlottesville, Virginia, USA

Harry Hochheiser
University of Pittsburgh
Pittsburgh, Pennsylvania, USA

Srini Venkatramanan
Biocomplexity Institute, University of Virginia
Charlottesville, Virginia, USA
srini@virginia.edu

Abstract

Epidemic forecasting has become an integral part of real-time infectious disease outbreak response. While collaborative ensembles composed of statistical and machine learning models have become the norm for real-time forecasting, standardized benchmark datasets for evaluating such methods are lacking. Further, there is limited understanding on performance of these methods for novel outbreaks with limited historical data. In this paper, we propose IDOBE, a curated collection of epidemiological time series focused on outbreak forecasting. IDOBE compiles from multiple data repositories spanning over a century of surveillance and across U.S. states and global locations. We perform derivative-based segmentation to generate over 10,000 outbreaks covering multiple outcomes such as cases and hospitalizations for 13 diseases. We consider a variety of information-theoretic and distributional measures to quantify the epidemiological diversity of the dataset. Finally, we perform multi-horizon short-term forecasting (1- to 4-week-ahead) through the progression of the outbreak using 11 baseline models and report on their performance. In addition to standard metrics such as NMSE and MAPE for point forecasts, we include probabilistic scoring rules such as Normalized Weighted Interval Score (NWIS) to quantify the performance. We find that MLP-based methods have the most robust performance, with statistical methods having a slight edge during the pre-peak phase. IDOBE dataset along with baselines are released publicly on https://github.com/NSSAC/IDOBE to enable standardized, reproducible benchmarking of outbreak forecasting methods.

Keywords

Forecasting, Benchmark, Epidemics, Timeseries, Machine Learning

1 Introduction

In recent years, epidemic forecasting has emerged as an active subdomain of computational epidemiology. Short-term forecasts of infectious disease activity have been adopted by various sub-national, national, and international agencies to guide outbreak response [36]. Agencies such as US Centers for Disease Control and Prevention (CDC) have established dedicated centers focused on improving the science, engineering, and translation of forecasting and outbreak analytics. Multi-model ensembles have been constituted to support both seasonal (e.g., Influenza) [40] and pandemic (e.g., COVID-19) [14] prediction efforts. Such efforts have been expanded to producing scenario-based projections to guide public health policy [8, 33]. Unlike projection models which mostly rely on mechanistic representations of underlying dynamics, forecast ensembles are constituted by a diverse collection of models [3, 43] including machine learning, statistical and mechanistic approaches, and take advantage of multiple data streams [1] including syndromic, clinical, and environmental surveillance [35], as well as internet-based indicators. Through this partnership, robust infrastructure has been developed to undertake such Hub-style efforts for future outbreaks [9, 30, 46].

While significant strides have been made in advancing real-time epidemic forecasting, there is a lack of standardized, multi-disease benchmark datasets for performance evaluation of existing models and ensembles. This is especially challenging in the context of operationalizing such models for a novel outbreak [1] either in a region with limited data availability (e.g., 2014-16 West African Ebola outbreak) or limited historical or seasonal data (e.g., COVID-19 in early 2020). While multiple real-time [7, 13] and retrospective [43] efforts have been undertaken, there is also need for standardized evaluation of methods outside the operational context. Most real-time efforts (outside COVID-19) have involved seasonal epidemics like Influenza, Dengue, and hence have leveraged the existence of historical data. Pre-trained epidemic models [29] trained across

[1] For our purposes, we will adopt CDC's definition of outbreak as a period with more disease cases than expected for a given time, within a specific location, and for a target population [19].

various regional outbreaks will be needed for rapid deployment and wider adoption.

1.1 Contributions

In this paper, we present IDOBE, an ecosystem for benchmarking models in the task of infectious disease outbreak forecasting. IDOBE comprises curated and preprocessed outbreaks drawn from diverse data repositories, along with a collection of baseline models and standardized evaluation metrics relevant for epidemic forecasting. Specifically:

- We preprocess epidemic time series datasets for **13 different diseases**, across **248 unique locations**, and outcomes such as outpatient visits, confirmed cases and hospitalizations. The dataset comprising **10799 outbreaks**, is compiled from existing disease data repositories such as Tycho [49], JHU-CSSE COVID-19 data repository [17], as well public health surveillance published by US CDC and the National Healthcare Safety Network (NHSN).
- We propose a suite of **information-theoretic** and **distributional measures** to characterize the diversity of outbreak trajectories contained in IDOBE. While some of these measures such as entropy [15] and permutation entropy [45] have been used in isolated studies, such a multi-dimensional characterization has not been performed before in the context of epidemic outbreak trajectories.
- We generate for multi-horizon short-term forecasting (1- to 4-week ahead) from **11 baseline models** across the progression of the outbreak. The baseline models span a variety of statistical (ARIMA, ETS), MLP-based (MLP, N-BEATS, N-HiTS), transformer-based (Informer,TFT), and RNN-based (RNN, GRU, LSTM, TCN) methods. In addition to producing point forecasts, the models are run with uncertainty quantification to produce probabilistic forecasts in the Hubverse [30] standard format consistent with existing forecasting Hubs.
- In addition to evaluating the point forecasts using standard metrics such as **MAPE, NMSE**, we also incorporate a **normalized** version of the **Weighted Interval Score (NWIS)** [10] for the probabilistic forecasts. We interpret the performance of baseline models across epidemiological context of the outbreak (pre-peak or post-peak) and forecast horizon as well as by disease.

Data and code availability. To ensure reproducibility and facilitate further exploration, we provide preprocessed datasets, trained baseline models, and scripts for extracting outbreak analytics and evaluation metrics through the public repository: https://github.com/NSSAC/IDOBE

1.2 Related Work

Collaborative forecasting "challenges" have been conducted for more than a decade under the Epidemic Prediction Initiative by US CDC, for targets ranging from seasonal influenza-like illness (ILI) forecasting (2013-now) [7, 40], Dengue (2015) [27], Chikungunya (2014) [16] and West Nile Virus [25]. Of these, ILI forecasting has received the most attention, with retrospective forecast performance evaluation conducted as part of the FluSight Network [43]. Similar efforts during the COVID-19 pandemic [13, 14] played a key role in influencing public policy, although model performance varied significantly across key epochs [34, 44]. While most of these efforts target seasonal, recurrent epidemics, IDOBE is designed around discrete outbreak episodes, mirroring novel or emerging pathogen scenarios.

Other efforts such as the M-competitions [37–39] have been undertaken in the forecasting community outside epidemiology. Recently, benchmarks such GIFT-Eval [5] have been developed for general time series forecasting tasks. Recently, Pre-trained [29] and Foundation models [28] have been developed and evaluated in the epidemiological context. A similar benchmarking framework to ours was envisioned in [47], although it was limited to COVID-19 forecasting in the US with simpler evaluation metrics and did not tackle the broader task of outbreak forecasting as outlined in this paper.

2 Methodology

2.1 Task definition

Consider an outbreak of disease d reported at location l for a particular disease outcome o (cases, deaths, or hospitalizations). Let i be the unique identification number of the outbreak, and let T_i be the duration of the outbreak. We denote the time series corresponding to the outbreak i as a vector $\mathbf{x}_z(0:T_i) = [x_z(0), x_z(1), \cdots, x_z(T_i)]$, where $x_z(\cdot)$ is the value of the outcome and $z = (d, l, o, i)$. In this work, we focus on simulating the task of short-term forecasting of outbreaks in real-time. Thus, given observations up to u, denoted as $\mathbf{x}_z(0:u)$, the goal is to forecast the values $\mathbf{x}_z(u+1:u+h)$, where h is the forecast horizon. In epidemic forecasting, similar to weather forecasting [22], predictive probability distribution of future values is the standardized format for reporting forecasts. Hence, the real-time forecasting task would involve learning a model $f_\Theta(k) : \mathbf{x}_z(0:u) \rightarrow P(x_z(u+k)|\mathbf{x}_z(0:u), \Theta)$, for $0 \le u \le T_i$, $0 \le k \le h$.

The model parameters Θ are typically learned from historically observed outbreaks. Typically, historical data consists of multiple outbreaks. In the subsequent sections, we discuss the process of extracting individual outbreaks from the complete time series.

2.2 Benchmark data curation

2.2.1 Available datasets. We collect disease-specific data from four different sources: Tycho [49], JHU-CSSE [17], US CDC, and NHSN. Each data source has data collected over different timelines, cover different diseases, temporal resolution (daily or weekly), and locations and the data format and nomenclature used also vary. Daily counts (only available for COVID-19 cases) are often dominated by reporting noise and day-of-the-week effects and can be smoothed out by aggregating daily data to weekly resolution. Moreover, many epidemic forecasting efforts require weekly forecasts [13, 40]. Accordingly, we align the temporal resolution across all datasets to be weekly, indexed by the MMWR week (Sunday-Saturday). Since COVID-19 confirmed cases are reported as daily counts, weekly count was obtained by summing the reported counts from Sunday-Saturday.

The Tycho dataset [49] contains weekly reports of 56 infectious diseases collected between 1888 and 2014 across various U.S. cities,

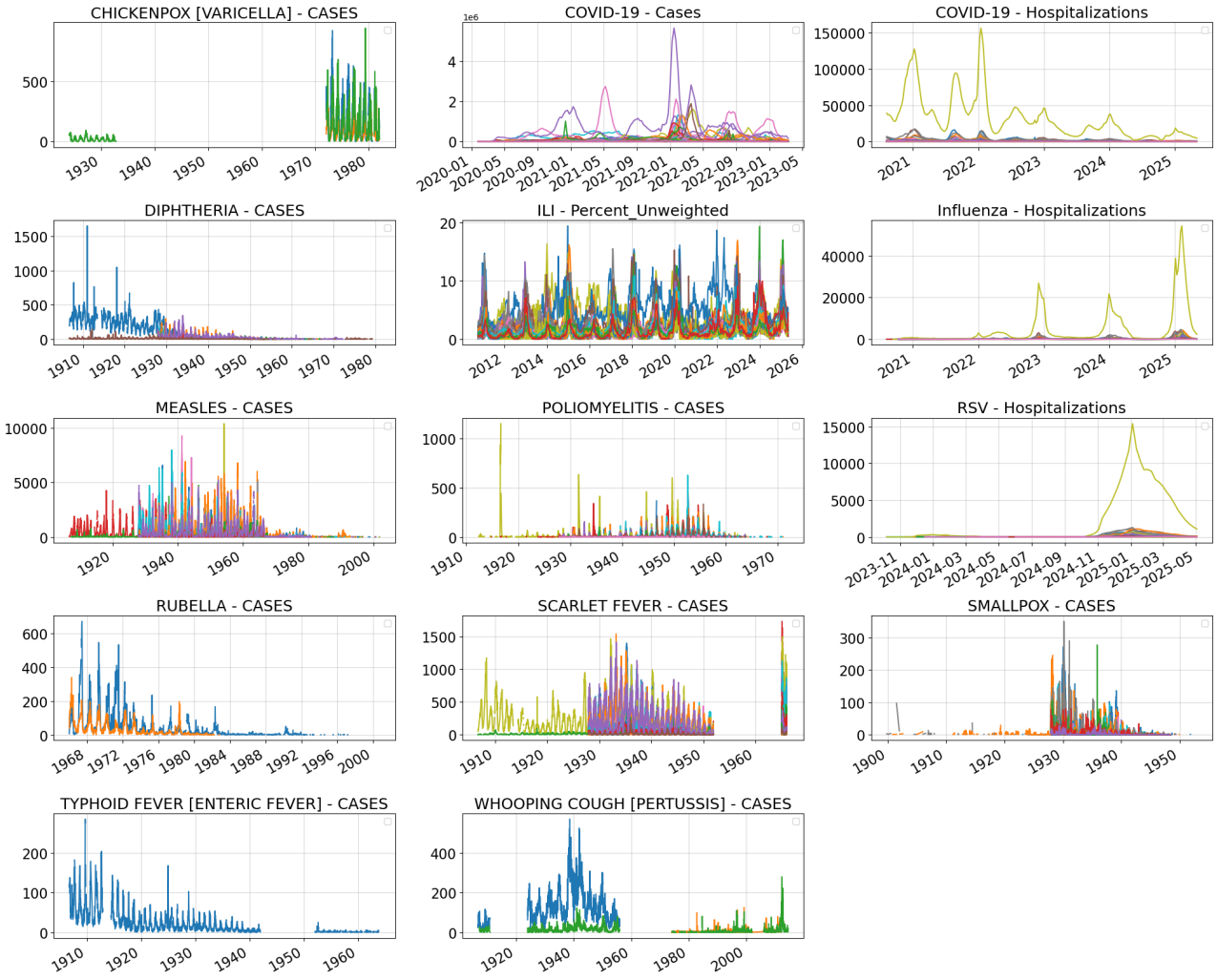


Figure 1: Timeseries corresponding to different diseases.

Table 1: Statistics of source datasets and outbreaks across different diseases

disease	event	# outbreaks	start date	end date	# years
CHICKENPOX [VARICELLA]	CASES	25	1972-02-19	1982-08-07	10
COVID-19	CASES	981	2020-03-21	2025-05-10	5
COVID-19	HOSPITALIZATIONS	473	2020-07-18	2025-05-31	5
DIPHTHERIA	CASES	1062	1927-12-03	1978-11-18	51
ILI	PERCENT UNWEIGHTED	867	2010-09-11	2025-05-31	15
INFLUENZA	HOSPITALIZATIONS	337	2020-10-24	2025-05-31	5
MEASLES	CASES	1517	1927-12-24	1997-05-17	69
POLIOMYELITIS	CASES	2442	1921-12-31	1971-08-21	50
RSV	HOSPITALIZATIONS	55	2023-11-11	2025-05-31	2
RUBELLA	CASES	84	1966-10-01	1997-03-08	30
SCARLET FEVER	CASES	1333	1927-11-05	1967-02-04	39
SMALLPOX	CASES	1324	1910-08-06	1949-05-07	39
TYPHOID FEVER [ENTERIC FEVER]	CASES	48	1928-04-21	1963-06-29	35
WHOOPING COUGH [PERTUSSIS]	CASES	251	1937-03-20	2015-05-09	78

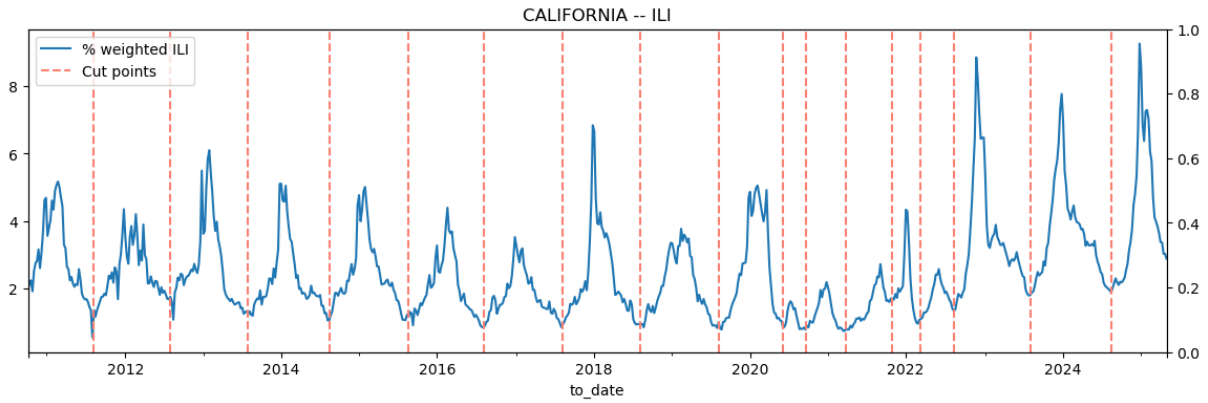


Figure 2: Example segmentation of timeseries into individual outbreaks (pink vertical dashed lines indicate the cutpoints).

counties, and states. However, no single disease was reported continuously throughout the entire interval. Among the Tycho time series, many contain substantial missing data. In order balance retention of long historical series with the need for sufficient observed data for model training, we exclude timeseries with significant proportion of missing values and no prominent outbreak segments. For the remaining timeseries, missing values are imputed using linear interpolation. After filtering, nine diseases from the Tycho dataset were retained.

The Johns Hopkins University (JHU) COVID-19 data repository [17] provides time series of reported COVID-19 cases globally, as well as across U.S. states and counties. At the global level, we include 201 locations, comprising countries and a few special administrative regions or events. At the U.S. level, we focus on all 50 states. Although the JHU dataset includes county-level data, it is often sparse and noisy due to low case counts; thus, we exclude it in this version. We plan to incorporate county-level data in a future release. We obtained the percentage of outpatient Influenza-Like Illness (ILI) visits data from CDC at the state level for United States, spanning fifteen seasons from 2010 to 2025 [2]. The National Healthcare Safety Network (NHSN) dataset [3] includes weekly records of new hospital admissions due to COVID-19, Influenza, and RSV across the 50 states of U.S. In total, we compile time series data for 13 diseases across the four data sources. The timeseries corresponding to different diseases are shown in Figure 1.

2.2.2 Outbreak Segmentation. Disease surveillance corresponding to every disease d, location l, outcome o consist of multiple outbreaks. We denote the historical reports as $X^o_{d,l}(t)$. In the benchmark dataset, we extract the individual outbreaks from each $X^o_{d,l}(t)$ using a segmentation function $S_\phi : X^o_{d,l}(t) \rightarrow \{x_{d,l,o}(t_n : t_n + T_n)\}_{n=0}^{N-1}$. Each individual outbreak is assigned a unique identifier i and stored in a dictionary $\mathcal{X} = \{\mathbf{x}_z\}_{z \in Z}$, where $z = (d, l, o, i)$ and Z is the set of all tuples.

We employ a derivative-based function as S_ϕ to obtain individual outbreaks. We use a Python toolbox EpidemicKabu [4][20], which is designed to identify epidemic waves by detecting peaks, valleys, and inflection points in time series data. We employ the wave identification functionality to segment a given time series into different waves. Wave detection involves (*i*) smoothing the timeseries using a Gaussian kernel, (*ii*) determining cut points in the smoothed first derivative where the first derivative crosses the x-axis from negative to positive, and (*iii*) selecting the detected cut points whose second derivative is less than a threshold. The output of EpidemicKabu consists of a list of cut points, where each segment between two consecutive cut points is interpreted as a potential outbreak. An example of the resulting outbreak segmentation is shown in Figure 2.

Such a segmentation is central to the task of outbreak forecasting for multiple reasons. Typical multi-wave patterns in epidemic time series emerge due to a combination of seasonal, geographical, demographical, and biological contexts. For example, influenza epidemics exhibit strong seasonality and are often characterized by staggered peaking dynamics of different age groups. COVID-19 pandemic saw the emergence of multiple variants that resulted in distinct waves. Further, non-pharmaceutical interventions and heterogeneous population mixing could result in geographic diversity of epidemic spread that may manifest as distinct "modes" in an aggregate epidemic curve. These characteristics are often muted in the case of a novel outbreak, and hence a robust forecasting framework must be capable of leveraging latent dynamics within an isolated outbreak trajectory.

We discard segments whose duration is less than 8 weeks or greater than 52 weeks for two reasons: (i) to minimize detection of brief spikes as outbreaks, (ii) to separate out extremely long multi-seasonal trends. Additionally, to ensure that sufficient context is available around each outbreak, we append four weeks of time series data both before the start and after the end of each identified segment, which result in overlapping segments across outbreaks. This provides sufficient context for models trained on fixed windows, and also helps avoid boundary effects in feature extraction. Metadata provided per outbreak includes region (state/country), time period, disease ontology, and indicators of sporadic/seasonal patterns. Table 2 provides the data dictionary provided to the user within the dataset.

[2]https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html
[3]https://www.cdc.gov/nhsn/psc/hospital-respiratory-dashboard.html
[4]https://pypi.org/project/EpidemicKabu/

Table 2: Data dictionary describing the metadata and structure of the dataset.

`unique_id`	An unique identification number for each outbreak
`disease`	Type of disease (COVID-19, influenza, Smallpox, etc.)
`location`	Name of the location (US states, countries, etc.)
`event`	Type of burden indicator (cases, hospitalizations, etc.)
`start_date`	Start date of the outbreak (end-of-week Saturday date)
`end_date`	End date of the outbreak (end-of-week Saturday date)
`duration`	Duration of the outbreak (in weeks)
$[0-59]$	Values observed for the particular outbreak for given week (counts, %)

2.3 Analytical measures

Since the dataset consists of diverse set of outbreaks collected over different periods, under different reporting strategies, and geographical locations, we analyze the characteristics of outbreaks using multiple statistical measures. These measures capture the uncertainty/noise and shapes of the different outbreaks.

Entropy analysis. Following the definition of *epidemic intensity* defined in [15], we compute the Shannon entropy of the incidence distribution of each outbreak, such that it is minimized when incidence is spread evenly across weeks and increases as incidence becomes more intensively focused in particular weeks. The incidence curve of an outbreak i is normalized, that is, $\bar{\mathbf{x}}_z(0:T_i) = \frac{1}{\sum_{t=0}^{T_i} x_z(t)} \mathbf{x}_z(0:T_i)$ to obtain the incidence distribution. The Shannon entropy is computed over the probability distribution.

Permutation Entropy analysis. We employ permutation entropy (PE) [6] to characterize the diversity of short-term ordinal patterns within individual outbreaks. PE is a model-free measure of uncertainty in a signal and has been used to quantify the uncertainty and predictability of epidemic time series [45]. In contrast to the Shannon entropy, PE does not consider the frequency of state changes but the frequency of permutation patterns (ordinal patterns) within a signal and is characterized by an embedding dimension (order) and a delay parameter. We consider embedding dimension of 3, referring to the number of consecutive data points in a time series that are grouped to form the embedding vector. This, in turn, determines the number of ordinal patterns that can be obtained (for an embedding dimension $d = 3$, the number of possible patterns would be $d! = 3!$). The delay parameter determines the temporal resolution at which the patterns are analyzed and we fix the delay to be 1 week (also known as no skip) in our analysis. Signals with high stochasticity (pure white noise) will likely have all patterns occur with equal frequency and are characterized by a high PE value. On the other hand, a perfectly periodic signal will typically have low entropy. We refer the readers to [6] for more details, insights, and examples of PE and its parameters.

Skewness and Kurtosis. These statistical measures obtained as the third and fourth moments of the incidence distribution help characterize the outbreak shape in terms of asymmetry (skewness) and tailedness/peakedness (kurtosis) relative to a normal curve.

2.4 Baseline methods

We evaluate four types of forecasting methods as baselines in this study: i). Statistical Methods, including ARIMA (Autoregressive Integrated Moving Average) and ETS (Exponential Smoothing) [21, 26]. ii). Recurrent neural network (RNN)-based methods, including GRU (Gated Recurrent Unit) [12], LSTM (Long Short Term Memory) [24], RNN (Recurrent Neural Network) [18], and TCN (Temporal Convolution Network) [31]. iii). Transformer-based methods, including Temporal Fusion Transformer (TFT) [32], and Informer [50]. iv) Multilayer Perceptron (MLP)-based methods: including vanilla MLP [23], Neural Basis Expansion Analysis for Time Series Forecasting (NBEATS) [42], and Neural hierarchical interpolation for time series forecasting (NHITS) [11]. All statistical and deep-learning model implementations are based on the unified and efficient forecasting framework provided by *Nixtla's* `StatsForecast` [21] and `NeuralForecast` [41] libraries. The motivation for using *Nixtla* is that it offers a unified API for a wide range of state-of-the-art statistical and deep learning models and robust support for hyperparameter tuning via Optuna or Ray. Its modular design and scalability make it well-suited for benchmarking diverse forecasting methods under consistent experimental settings.

All the models are trained to produce probabilistic forecasts. Nixtla's ARIMA and ETS produce uncertainty through parametric predictive distributions implied by the fitted statistical models. Forecast variances are derived analytically from estimated model residuals assuming Gaussian errors and propagated across horizons to obtain prediction intervals and quantiles. Nixtla's deep learning models generate uncertainty through quantile regression, training neural networks with multi-quantile (pinball) losses to predict multiple forecast quantiles without assuming an explicit data-generating distribution. We set the number of quantiles to 23 consistent with the recommendations from the FluSight and COVID-19 forecast hubs [13, 40].

2.4.1 Model Training/Fitting and Testing. To accommodate the different classes of baselines, we adopt a model-class-specific training/fitting strategies. Since all models are intended for real-time deployment (Section 2.1), they are designed to operate under an expanding window setting (see timeseries cross-validation technique in [26]). Specifically, as an outbreak progresses, models have access to an expanding set of observations $\mathbf{x}_z(0:u)$, $7 \leq u \leq T_i - h, \forall z$. We impose a minimum window of eight and a maximum of $T_i - h$ for training/fitting of the models.

In the case of statistical models (ARIMA, ETS), for a given z, a model is fit on each of the expanding-window of observations $\mathbf{x}_z(0:u)$, $7 \leq u \leq T_i - h$. The testing is performed by generating out-of-sample forecasts from the fitted model for horizons $(u+1 : u+h)$, $7 \leq u \leq T_i - h$ and evaluating the forecasts against the observed values. We employ the `AutoARIMA` model provided by Nixtla to automatically select the optimal ARIMA parameters.

For the remaining model classes that require explicit training, we adopt a unified training/validation/testing strategy. Rather than

splitting data along the temporal dimension, we partition the dataset across outbreaks. We shuffle the unique_ids and split them into training/validation/testing sets in a 60%/20%/20% fashion, respectively. Consequently, during testing, models are evaluated on completely unseen outbreaks.

During training, we tune key hyperparamters and select the optimal model parameters based on the model's performance on the validation set. We employed the hyperparameter optimization framework Optuna [4] to determine the optimal set of hyperparameters. Following are the set of key tunable hyperparameters along with the search space:

- `input_size`, which controls the length of the historical look-back window used as input to the model, sampled as an integer between $[8, \texttt{max\_input\_size}]$.
- `learning_rate`, which specifies the optimizer step size during model training and is sampled on a logarithmic scale between 10^{-4} and 10^{-2}.
- `batch_size`, which specifies the number of samples processed per training optimization step and is selected from the set $\{16, 32, 64\}$.

Forecasting on the full test set requires generating $48 \times 4 = 192$ forecasts for each of the 2000 outbreaks and is time consuming. To enable efficient evaluation, we randomly sample minibatches of 100 outbreaks from the test set and compute the forecast performance on each minibatch. This procedure is repeated multiple times, and final performance metrics are obtained by averaging the performance across minibatches.

2.5 Forecast Evaluation Metrics

2.5.1 Point Forecasts. Although our baseline models are designed to provide probabilistic forecasts, to lower the barrier for entry for new models, we also include evaluation metrics for point forecasts. Specifically, we include normalized versions of error metrics to be scale-agnostic across the outbreaks. We consider mean absolute percentage error (MAPE) and normalized mean square error (NMSE) [48] obtained for each forecast target, and averaged across outbreaks, timepoints, and horizons. For the baseline model forecasts, we assumed the values corresponding to the median (quantile level = 0.5) as the point forecasts.

2.5.2 Probabilistic Forecasts. In order to compare the forecast quantiles of the different models, we use the Weighted Interval Score (WIS), the de facto standard in epidemiological forecasting community for probabilistic forecast evaluation [10]:

$$WIS_{\alpha_{0:k}}(F, y) = \frac{1}{K+0.5} \sum_{k=0}^{K} \frac{\alpha_k}{2}(u_k - l_k) + \frac{2}{\alpha_k}(l_k - y)\mathbb{1}(y < l_k) + \frac{2}{\alpha_k}(y - u_k)\mathbb{1}(y > u_k) \quad (1)$$

where y is the observed value (ground truth case count corresponding to a week) for a given location and date, F is the forecast defined in terms of the median m, upper quantiles u_k and lower quantiles l_k of the predictive distribution, respectively. K is the number of intervals considered, which in our case $K = 11$. Since WIS is scale dependent, we divide the obtained WIS for a given target by its ground truth value, hence yielding a normalized weighted interval score (NWIS).

3 Results

3.1 Dataset characteristics

Before describing the insights obtained from the analytical measures, we note that as seen in Table 1, the number of outbreaks vary widely across diseases, depending on the historical trends as well as quality of data capture. For instance, historical vaccine-preventable disease such as poliomyelitis and measles contribute the largest number of outbreaks. Even though spanning fewer years of data capture, due to the global nature of COVID-19 pandemic and seasonal patterns in influenza, they contribute sizeable number of outbreaks as well.

3.2 Analytical measures

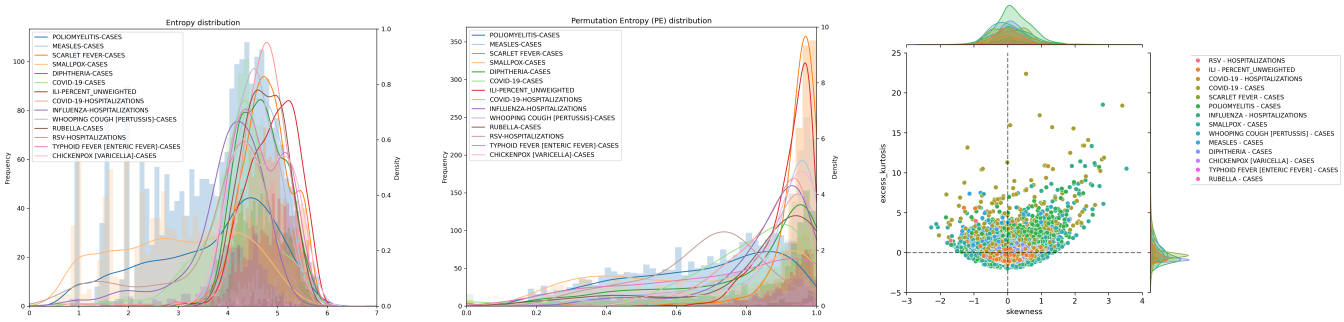
Figure 3a shows the distribution of entropies across outbreaks obtained per (disease, outcome) tuple. We observe an universal mode centered around 5, with heterogeneity across diseases. Diseases such as poliomyelitis and smaller have broader entropy distributions, indicating the presence of both sharp (i.e., low entropy) and flatter outbreaks. From Figure 3b, we note that most outbreaks have high entropy among ordinal patterns of order 3, thus signaling significant noisiness. Certain outbreaks such as those of RSV hospitalizations seem to have lower PE thus hinting at better predictability [45]. Finally, as seen in Figure 3c, most of the outbreaks have negative excess kurtosis (i.e., platykurtic) and positive skew compared to the normal curve. While negative kurtosis confirms the presence of flatter outbreaks, positive skewness indicates the typical "left-skewed" nature of epidemic curves with steep inclines before peak and slower declines post-peak. In addition to limited training data, such a characteristic could also result in lower predictability for the early phases for a novel outbreak.

3.3 Baseline performance

Table 3 shows the performance of the 11 baseline models across different forecast horizons (1- to 4-week ahead) as well as combined performance. We note that performance degrades quickly across horizons. For shorter horizons, both statistical (ETS) and transformer based (TFT) methods perform best. For longer horizones, MLP-based methods have better performance, with MLP performing best for all error metrics for the 4-week ahead.

Further, we consider post- and pre-peak performance since epidemiological these might vary in relevance for policymakers [48]. We observe that the statistical models (ETS) perform best in pre-peak time points, where as transformer-based methods (TFT) perform better for post-peak time points. We also observe that MLP-based methods performance consistently well across both pre- and post-peak phases. This phase-dependent behavior underscores the importance of phase-specific model training and selection [2].

Model performance also vary across diseases and outcomes. As seen in Figure A.1, all baseline models seem to have poorer performance for poliomyelitis and smallpox cases, with better performance for ILI and Influenza hospitalizations. Average model performance across forecast horizons is shown in Figure A.2, with



(a) Entropy distribution **(b) Permutation Entropy distribution** **(c) Skewness and Excess Kurtosis**

Figure 3: Analytical measures on the IDOBE dataset

more gradual degradation seen for LSTM and TFT, where statistical models like ARIMA and ETS show a more rapid decline in performance with forecast horizon.

4 Discussion

In this paper, we presented IDOBE, a novel infectious disease outbreak forecasting benchmark ecosystem, comprising preprocessed outbreaks, baseline models, and evaluation metrics. In addition to characterizing the epidemic outbreaks through various analytical measures, we summarize the performance of baseline models within the epidemiological context. Note that the core premise of this work is the need for and ability to segment epidemic surveillance into single-wave outbreaks for benchmarking purposes. We acknowledge that such an approach inherently neglects multi-wave dynamics that may arise from policy changes, spatial heterogeneity and viral evolution. IDOBE, in its current form, is best seen as a benchmark for single-outbreak short-term forecasting and fills an obvious gap in methods development for effective infectious disease response.

We also focus on univariate forecasting at the level of single location, and disease outcomes. Spatial coupling and models that leverage relationship among multivariate signals have found success in real-time epidemic forecasting. Further, we have not considered mechanistic constraints (e.g., population sizes, disease models) have not been exploited in the current baseline models. We intend to expand IDOBE in future versions to include more sophisticated baseline methods as well as additional epidemiological relevant forecast targets (e.g., peak intensity, total duration). Finally, we also envision incorporating epidemic simulators to produce plausible synthetic outbreaks to augment such benchmark datasets.

5 Data and model availability

IDOBE is openly available at https://github.com/NSSAC/IDOBE. The repository includes of (i) outbreak data comprising over 10,000 outbreak timeseries across multiple diseases, (ii) tools for data preprocessing, (iii) scripts for extracting analytical measures to analyze outbreaks, (iv) a suite of trained baseline forecasting models, (v) probabilistic forecast formatting and evaluation scripts.

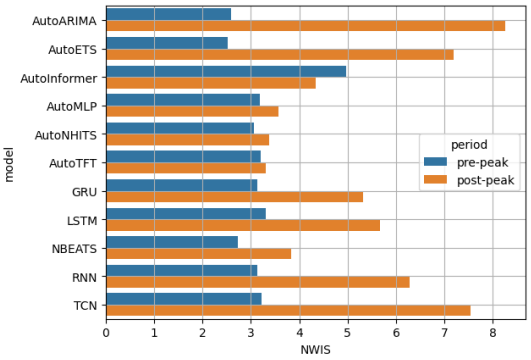


Figure 4: Forecast performance by NWIS pre- and post-outbreak peak date.

Table 3: Summary of forecast performance across models and horizons. Best (lowest) value per row is boldfaced.

Horizon	Metric	Statistical		MLP			Transformer		RNN			
		ARIMA	ETS	MLP	NBEATS	NHITS	Informer	TFT	GRU	LSTM	RNN	TCN
4-week	NWIS	6.4	5.7	3.4	3.5	**3.3**	4.5	3.3	4.6	4.9	5.3	6.1
	NMSE	1.6	**1.3**	1.5	1.4	1.4	2.5	1.4	1.6	1.6	1.5	1.5
	MAPE	113.3	76.2	67.3	65.8	62.9	89.7	**61.3**	72.4	74.6	80.4	115.9
1st wk	NWIS	4.0	4.1	3.4	2.7	2.8	4.5	**2.7**	4.3	4.5	4.8	4.8
	NMSE	1.2	**1.1**	1.6	1.3	1.4	2.6	1.2	1.5	1.6	1.5	1.4
	MAPE	67.1	56.1	65.3	51.5	53.7	87.3	**49.5**	68.1	68.2	76.7	89.7
2nd wk	NWIS	5.6	5.2	3.4	3.2	**3.1**	4.5	3.1	4.5	4.7	5.0	5.8
	NMSE	1.4	**1.3**	1.5	1.3	1.3	2.5	1.4	1.6	1.6	1.5	1.5
	MAPE	95.3	69.0	68.1	60.9	58.9	88.5	**58.6**	70.7	72.2	78.0	110.0
3rd wk	NWIS	7.3	6.2	3.5	3.7	**3.4**	4.6	3.5	4.7	5.0	5.4	6.8
	NMSE	1.8	1.5	1.4	**1.4**	1.5	2.5	1.5	1.6	1.6	1.5	1.6
	MAPE	130.4	82.6	67.7	70.6	66.6	91.2	**65.5**	74.0	76.6	81.1	122.4
4th wk	NWIS	9.1	7.4	**3.5**	4.3	3.8	4.7	3.9	5.0	5.4	5.8	7.4
	NMSE	2.1	1.6	**1.3**	1.4	1.6	2.5	1.5	1.5	1.6	1.5	1.6
	MAPE	166.2	99.9	**68.1**	81.9	73.5	92.1	72.8	77.3	82.1	86.4	144.9

References

[1] Aniruddha Adiga, Gursharn Kaur, Benjamin Hurt, Lijing Wang, Przemyslaw Porebski, Srinivasan Venkatramanan, Bryan Lewis, and Madhav Marathe. Enhancing covid-19 ensemble forecasting model performance using auxiliary data sources. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1594–1603. IEEE, 2022.

[2] Aniruddha Adiga, Gursharn Kaur, Benjamin Hurt, Lijing Wang, Przemyslaw Porebski, Srinivasan Venkatramanan, Bryan Lewis, and Madhav Marathe. Enhancing covid-19 ensemble forecasting model performance using auxiliary data sources. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1594–1603, 2022.

[3] Aniruddha Adiga, Lijing Wang, Benjamin Hurt, Akhil Peddireddy, Przemyslaw Porebski, Srinivasan Venkatramanan, Bryan Leroy Lewis, and Madhav Marathe. All models are useful: Bayesian ensembling for robust high resolution covid-19 forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2505–2513, 2021.

[4] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[5] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

[6] Christoph Bandt and Bernd Pompe. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102, 2002.

[7] Matthew Biggerstaff, David Alper, Mark Dredze, Spencer Fox, Isaac Chun-Hai Fung, Kyle S Hickmann, Bryan Lewis, Roni Rosenfeld, Jeffrey Shaman, Ming-Hsiang Tsou, et al. Results from the centers for disease control and prevention's predict the 2013–2014 influenza season challenge. *BMC infectious diseases*, 16:1–10, 2016.

[8] Rebecca K Borchering, Jessica M Healy, Betsy L Cadwell, Michael A Johansson, Rachel B Slayton, Megan Wallace, and Matthew Biggerstaff. Public health impact of the us scenario modeling hub. *Epidemics*, 44:100705, 2023.

[9] Nikos I Bosse, Hugo Gruson, Anne Cori, Edwin van Leeuwen, Sebastian Funk, and Sam Abbott. Evaluating forecasts with scoringutils in r. *arXiv preprint arXiv:2205.07090*, 2022.

[10] Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *arXiv preprint arXiv:2005.12881*, 2020.

[11] Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico García, Gonzalo Mena, and Artur Dubrawski. NHITS: Neural hierarchical interpolation for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6989–6997, 2023.

[12] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[13] Estee Y Cramer, Yuxin Huang, Yijin Wang, Evan L Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Katie House, et al. The united states covid-19 forecast hub dataset. *medRxiv*, 2021.

[14] Estee Y Cramer, Evan L Ray, Velma K Lopez, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Tilmann Gneiting, Katie H House, Yuxin Huang, et al. 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, 2022.

[15] Benjamin D Dalziel, Stephen Kissler, Julia R Gog, Cecile Viboud, Ottar N Bjørnstad, C Jessica E Metcalf, and Bryan T Grenfell. Urbanization and humidity shape the intensity of influenza epidemics in us cities. *Science*, 362(6410):75–79, 2018.

[16] Sara Y Del Valle, Benjamin H McMahon, Jason Asher, Richard Hatchett, Joceline C Lega, Heidi E Brown, Mark E Leany, Yannis Pantazis, David J Roberts, Sean Moore, et al. Summary results of the 2014-2015 darpa chikungunya challenge. *BMC infectious diseases*, 18:1–14, 2018.

[17] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[18] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[19] Centers for Disease Control and Prevention. Outbreak and case definitions. https://www.cdc.gov/urdo/php/surveillance/outbreak-case-definitions.html. Accessed: 2025-05-14.

[20] Lina Marcela Ruiz Galvis, Anderson Alexis Ruales Barbosa, Oscar Ignacio Mendoza Cardozo, Noël Christopher Barengo, Jose L Peñalvo, and Paula Andrea Diaz Valencia. Epidemickabu a new method to identify epidemic waves and their peaks and valleys. *medRxiv*, pages 2024–03, 2024.

[21] Azul Garza, Max Mergenthaler Canseco, Cristian Challú, and Kin G. Olivares. StatsForecast: Lightning fast forecasting with statistical and econometric models. PyCon Salt Lake City, Utah, US 2022, 2022.

[22] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 2007.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[25] Karen M Holcomb, Sarabeth Mathis, J Erin Staples, Marc Fischer, Christopher M Barker, Charles B Beard, Randall J Nett, Alexander C Keyel, Matteo Marcantonio, Marissa L Childs, et al. Evaluation of an open forecasting challenge to assess skill of west nile virus neuroinvasive disease prediction. *Parasites & Vectors*, 16(1):11, 2023.

[26] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[27] Michael A Johansson, Karyn M Apfeldorf, Scott Dobson, Jason Devita, Anna L Buczak, Benjamin Baugher, Linda J Moniz, Thomas Bagley, Steven M Babin, Erhan Guven, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274, 2019.

[28] Suprabhath Kalahasti, Benjamin Faucher, Boxuan Wang, Claudio Ascione, Ricardo Carbajal, Maxime Enault, Christophe Vincent Cassis, Titouan Launay, Caroline Guerrisi, Pierre-Yves Boëlle, et al. Foundation time series models for forecasting and policy evaluation in infectious disease epidemics. *medRxiv*, pages 2025–02, 2025.

[29] Harshavardhan Kamarthi and B Aditya Prakash. Pems: Pre-trained epidemic time-series models. *arXiv preprint arXiv:2311.07841*, 2023.

[30] Melissa Kerr, Rebecca Borchering, Alvaro Castro Rivadeneira, Lucie Contamin, Sebastian Funk, Harry Hochheiser, Emily Howerton, Anna Krystalli, Li Shandross, Nicholas G Reich, et al. Coordinating collaborative infectious disease modeling projects with the hubverse. *medRxiv: the preprint server for health sciences*, 2025.

[31] Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 156–165, 2017.

[32] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

[33] Sara L Loo, Emily Howerton, Lucie Contamin, Claire P Smith, Rebecca K Borchering, Luke C Mullany, Samantha Bents, Erica Carcelen, Sung-mok Jung, Tiffany Bogich, et al. The us covid-19 and influenza scenario modeling hubs: delivering long-term projections to guide policy. *Epidemics*, 46:100738, 2024.

[34] Velma K Lopez, Estee Y Cramer, Robert Pagano, John M Drake, Eamon B O'Dea, Madeline Adee, Turgay Ayer, Jagpreet Chhatwal, Ozden O Dalgic, Mary A Ladd, et al. Challenges of covid-19 case forecasting in the us, 2020–2021. *PLoS computational biology*, 20(5):e1011200, 2024.

[35] Raimundo Seguí López-Peñalver, Rubén Cañas-Cañas, Jorge Casaña-Mohedo, José Vicente Benavent-Cervera, Julio Fernández-Garrido, Raúl Juárez-Vela, Ana Pellín-Carcelén, Vicente Gea-Caballero, and Vicente Andreu-Fernández. Predictive potential of SARS-Cov-2 RNA concentration in wastewater to assess the dynamics of COVID-19 clinical outcomes and infections. *Science of The Total Environment*, 886:163935, 2023.

[36] Chelsea S Lutz, Mimi P Huynh, Monica Schroeder, Sophia Anyatonwu, F Scott Dahlgren, Gregory Danyluk, Danielle Fernandez, Sharon K Greene, Nodar Kipshidze, Leann Liu, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, 19(1):1659, 2019.

[37] Spyros Makridakis and Michele Hibon. The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476, 2000.

[38] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020.

[39] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4):1325–1336, 2022.

[40] Sarabeth M Mathis, Alexander E Webber, Tomás M León, Erin L Murray, Monica Sun, Lauren A White, Logan C Brooks, Alden Green, Addison J Hu, Roni Rosenfeld, et al. Evaluation of flusight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations. *Nature communications*, 15(1):6289, 2024.

[41] Kin G. Olivares, Cristian Challú, Federico Garza, Max Mergenthaler Canseco, and Artur Dubrawski. NeuralForecast: User friendly state-of-the-art neural forecasting models. PyCon Salt Lake City, Utah, US 2022, 2022.

[42] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *International Conference on Learning Representations*, 2020.

[43] Nicholas G Reich, Logan C Brooks, Spencer J Fox, Sasikiran Kandula, Craig J McGowan, Evan Moore, Dave Osthus, Evan L Ray, Abhinav Tushar, Teresa K Yamana, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019.

[44] Roni Rosenfeld and Ryan J Tibshirani. Epidemic tracking and forecasting: Lessons learned from a tumultuous year. *Proceedings of the National Academy of Sciences*, 118(51):e2111456118, 2021.

[45] Samuel V Scarpino and Giovanni Petri. On the predictability of infectious disease outbreaks. *Nature communications*, 10(1):898, 2019.

[46] Li Shandross, Emily Howerton, Lucie Contamin, Harry Hochheiser, Anna Krystalli, Nicholas G Reich, Evan L Ray, et al. Multi-model ensembles in infectious disease and public health: Methods, interpretation, and implementation in r. *medRxiv*, pages 2024–06, 2025.

[47] Ajitesh Srivastava, Tianjian Xu, and Viktor K Prasanna. The epibench platform to propel ai/ml-based epidemic forecasting: A prototype demonstration reaching human expert-level performance. In *International Workshop on Health Intelligence*, pages 165–179. Springer, 2021.

[48] Farzaneh Sadat Tabataba, Prithwish Chakraborty, Naren Ramakrishnan, Srinivasan Venkatramanan, Jiangzhuo Chen, Bryan Lewis, and Madhav Marathe. A framework for evaluating epidemic forecasts. *BMC infectious diseases*, 17(1):345, 2017.

[49] Willem G van Panhuis, Anne Cross, and Donald S Burke. Project tycho 2.0: a repository to improve the integration and reuse of data for global population health. *Journal of the American Medical Informatics Association*, 25(12):1608–1617, 2018.

[50] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021.

A Disease- and Outcome-Specific Baseline Performance

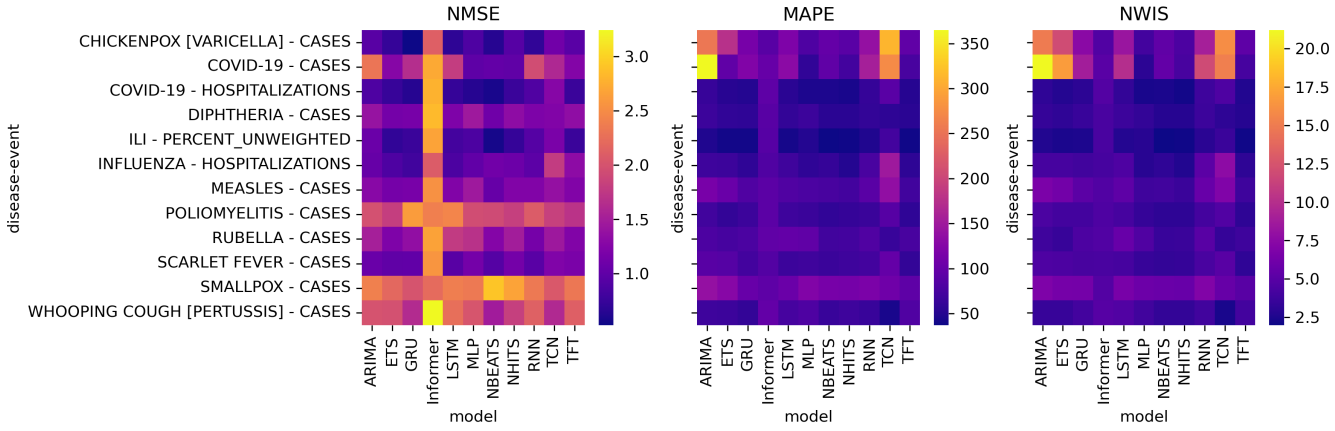


Figure A.1: Disease-specific performance across models by different error metrics

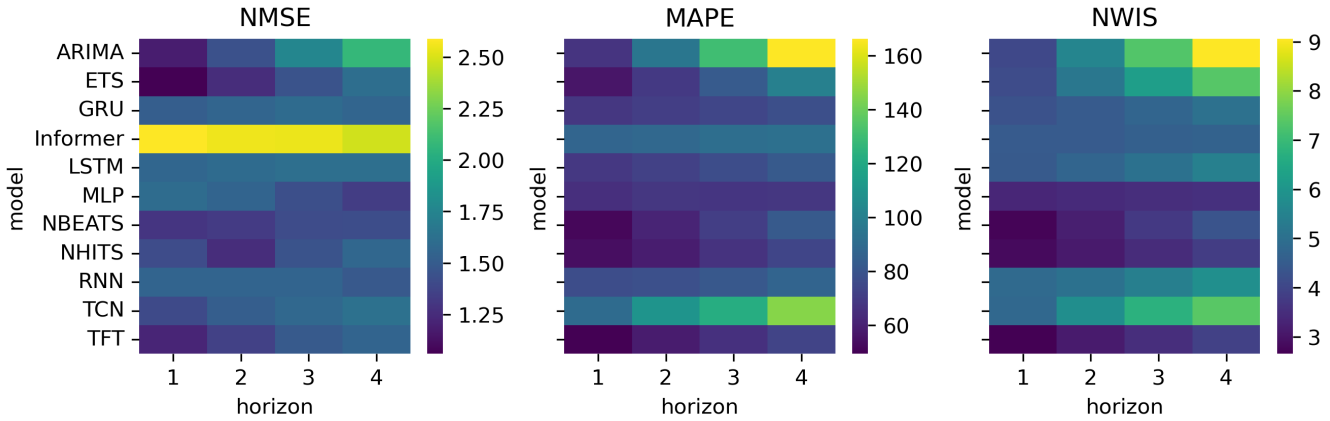


Figure A.2: Average model performance across forecast horizons.