Performance evaluation of RespiCast ensemble forecasts for primary care syndromic indicators of viral respiratory disease in Europe during the 2023/24 winter season

Nicolò Gozzi[1,*], Corrado Gioannini[1], Paolo Milano[1], Ivan Vismara[1], Luca Rossi[1], Marco Quaggiotto[2,1], Stefania Fiandrino[1,3], Daniela Paolotti[1], Alessandro Vespignani[4,1], Helen Johnson[5], Rok Grah[5], Sebastian Funk[6], Katharine Sherratt[6], Niel Hens[7,8], Steven Abrams[7,9], Maikel Bosschaert[7,10], Francesco Celino[11], Lorenzo Zino[11], Alessandro Rizzo[11], Sasikiran Kandula[12], Birgitte Freiesleben de Blasio[12,13], Atte Aalto[14], Daniele Proverbio[15], Giulia Giordano[15], Jorge Gonçalves[14], Yuhan Li[16], Nicola Perra[16,17], Ajibola Omokanye[18], Leah J. Martin[18], Rene Niehus[18], Jose Canevari[18], Eva Bons[18]

[1]ISI Foundation, Turin, Italy
[2]Department of Design, Politecnico di Milano
[3]Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome, Rome, Italy
[4]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA USA
[5]Safinea
[6]Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom
[7]Interuniversity Institute for Biostatistics and statistical Bioinformatics, Data Science Institute, UHasselt, Diepenbeek, Belgium
[8]Centre for Health Economics Research and Modelling Infectious Diseases, Vaccine and Infectious Disease Institute, Antwerp, Belgium
[9]Global Health Institute, Department of Family Medicine and Population Health, University of Antwerp, Antwerp, Belgium
[10]Julius Center for Health Sciences and Primary Care University Medical Center Utrecht, Utrecht University
[11]Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy
[12]Norwegian Institute of Public Health, Oslo, Norway
[13]University of Oslo, Oslo, Norway
[14]Luxembourg Centre for Systems Biomedicine, University of Luxembourg
[15]Department of Industrial Engineering, University of Trento, via Sommarive 9, Trento 38123, Italy
[16]School of Mathematical Sciences, Queen Mary University of London, UK
[17]The Alan Turing Institute, London, UK
[18]European Centre for Disease Prevention and Control

* nicolo.gozzi@isi.it

Abstract

During the 2023/24 winter season, the European Centre for Disease Prevention and Control (ECDC) launched RespiCast, the first European Respiratory Diseases Forecasting Hub, to provide probabilistic forecasts for influenza-like illness (ILI) and acute respiratory infection (ARI) incidence across 27 European countries. Spanning 20 forecasting rounds, RespiCast collected one- to four-week-ahead forecasts from 14 models for ILI and 10 models for ARI, contributed by different international teams, and combined them into an ensemble. Our analysis shows that the ensemble consistently outperformed the baseline model (a naive, uninformed model that projects the last observed value forward) in over 90% of countries for both ILI and ARI incidence when evaluated using the Weighted Interval Score (WIS), with the ensemble outperforming individual models in most forecasting rounds. Analysis of ensemble prediction coverage (proportion of times that the observed values fall within the specified prediction intervals) indicated that forecast prediction intervals were reliable in the short term, though an overconfidence trend (i.e. prediction intervals that are too narrow) was observed at longer forecasting horizons. The performance of the ensemble declined in certain weeks, likely due to reduced participation from modeling teams, rapid shifts in epidemic dynamics during specific phases, higher data noise, and reporting delays. Forecast scores varied across countries, with some exhibiting consistently higher errors than others. Overall, the findings highlight the strengths of ensemble forecasting in improving the accuracy and reliability of epidemiological forecasts while also identifying areas for improvement, such as managing overconfidence and addressing variability in performance across countries and over time.

Introduction

Infectious disease forecasting can play a critical role in informing health authorities about the dynamics of epidemic outbreaks, guiding the timely implementation of public health measures and effective communication strategies, as well as optimizing resource allocation [1], [2]. In this context, collaborative forecasting hubs have emerged as an effective approach to the challenges linked to infectious disease forecasting [3], [4]. In particular they address the characterization of uncertainty in disease modelling by integrating different approaches which capture different aspects of epidemic dynamics. Since no single model is likely to account for all the biological, behavioural, and environmental factors influencing an outbreak, combining multiple model outputs allows for a more comprehensive representation of possible epidemic trajectories.

The forecasting hub approach brings together diverse modelling teams with the ultimate goal of creating ensemble forecasts (i.e., combined predictions from multiple models) that, in past efforts, have consistently outperformed the predictions of individual models across multiple scoring criteria [5], [6], [7]. Besides their operational focus, these hubs also serve as collaborative research platforms to test novel forecasting approaches, develop new model evaluation and ensembling techniques, and ultimately promote continuous scientific advancement through shared insights.

During the COVID-19 pandemic, the European Centre for Disease Prevention and Control (ECDC) introduced the European COVID-19 Forecasting Hub that supported the EU's management of the health crisis [5]. Since 2021, this hub has provided real-time forecasts of cases, deaths, and hospitalizations related to COVID-19, with the aim of supporting public-health decision making. As we moved out of the acute phases of the COVID-19 pandemic, Europe shifted toward an integrated surveillance system for respiratory pathogens, now monitoring multiple viruses with a unified framework. A concrete example of this is the European Respiratory Virus Surveillance Summary (ERVISS), a joint effort between ECDC and the WHO Regional Office for Europe which provides weekly integrated summaries of the epidemiological and virological situation for influenza, respiratory syncytial virus (RSV), and SARS-CoV-2 in Europe, based on national surveillance data submitted to The European Surveillance System (TESSy) [8]. This approach to integrated respiratory pathogen surveillance calls for a similarly integrated strategy in forecasting; namely, the establishment of a unified European forecasting hub for viral respiratory diseases, capable of producing forecasts across multiple pathogens.

In this study, we present the structure and results of RespiCast, a new European respiratory diseases forecasting hub, launched by ECDC during the 2023/24 winter season. While both epidemiological and pathogen-specific, virological indicator targets are featured in ERVISS, as a starting point in the first season, RespiCast focused on primary care *epidemiological* indicators—specifically influenza-like illness (ILI) and acute respiratory infection (ARI)—to provide broad representation of viral respiratory activity in the community. This approach provides a scalable foundation, with scope to incorporate additional indicators in future iterations.

RespiCast conducted over 20 forecasting rounds in its first season, aggregating predictions from multiple modelling teams for ILI and ARI incidence across 27 European countries. We evaluate the performance of the ensemble forecasts generated by RespiCast by comparing them to individual models and a naive baseline and examine how forecast accuracy varies over time and across different geographical contexts.

Methods

Description of the RespiCast forecasting hub

During the 2023/24 respiratory virus season, modelling teams worldwide were invited to participate in the inaugural RespiCast forecasting challenge, which took place over 20 rounds (i.e., forecast submission cycles), spanning from 2023/12/20 to 2024/05/01 (more details on contributing teams in the Supplementary Information). Data were typically updated on Fridays, covering the previous week, and teams were required to submit their forecasts by the following Wednesday. These forecasts were then combined into an ensemble and published on the RespiCast website [9] every Thursday. The forecasting horizons ranged from 1 to 4 weeks ahead. Given the timing of the data updates, the forecasts published on Thursday covered the preceding week (for which public data had not yet been consolidated), the current week, and the two subsequent weeks. More formally, let r denote a forecasting round and t_r the reference time for that round, corresponding to the most recent epidemiological week for which data was available at the time of submission. For each round r , participating teams submitted probabilistic forecasts for multiple horizons $h \in \{1,2,3,4\}$ weeks ahead. The target time for a forecast at horizon h is then $t_r + h$, and the corresponding observed outcome (e.g., reported ILI incidence) is denoted by $y_{t_r+h}^{(c)}$ for country c. All forecast distributions $F_{r,h}^{(c)}$ are constructed to predict these future outcomes. Only probabilistic forecasts in the form of quantiles were accepted, and teams were free to submit forecasts for one or multiple countries. Full details on the submission format and requirements are provided in the GitHub RespiCast repository [10], [11].

Forecasting targets

Teams were asked to submit forecasts on both ILI and ARI incidence, expressed as cases per 100,000 population in all countries except for Cyprus, Luxembourg, Malta, and Finland where rates are expressed per 100,000 consultations. Both are syndromic indicators typically estimated through sentinel primary care surveillance in most countries. Specific case definitions can be found in Ref. [12]. Historical data from past seasons were made available to the modelling teams, and participants were also allowed to use additional data sources of their choice to inform their models (e.g., serological data, social contact data). ILI incidence was available for 27 European countries, while ARI incidence was available for 18 countries. Data for all countries was sourced from the European Respiratory Virus Surveillance Summary

(ERVISS) [8], except for England, Scotland, Wales, Northern Ireland, and Switzerland, for which we considered data from WHO FluID [13].

Multiple pathogens contribute to ILI and ARI incidence rates. However, they differ in their sensitivity and specificity for representing various respiratory pathogens, which supports the use of both as forecasting targets to better track the progression of respiratory diseases in Europe. The ARI case definition, for instance, does not include high temperature, making it generally more sensitive to a wider range of pathogens like COVID-19 and RSV. However, this more inclusive case definition comes at the cost of lower specificity compared to ILI, particularly for influenza [12].

Ensemble and baseline forecasting model

RespiCast combines multi-model projections into an ensemble forecast considering a median-across-quantiles approach [14]. For each country c, round r, and horizon h, let $F_{m,r,h}^{(c)}(q)$ denote the predictive quantile of level q submitted by model m. The ensemble forecast is defined as the quantile-wise median across models:

$$F_{ens,\,r,\,h}^{(c)}(q) \;= median_{\,m}\left\{F_{m,\,r,\,h}^{(c)}(q)\right\},$$

where $F_{ens,\,r,\,h}^{(c)}(q)$ is the ensemble's predictive quantile at level q.

The baseline is a naive, largely uninformed model which serves as a neutral benchmark for other models. Across all time horizons, the baseline model predicts the last data point within the calibration period as its median, while the prediction intervals are based on past data increments. The baseline forecast will thus always be a flat line with symmetric prediction intervals reflecting country-specific historical variation in the data. More precisely, for each country c, the baseline forecast for round r and horizon h is constructed by extrapolating forward from the most recently observed data point $y_{t_r}^{(c)}$, where t_r is the reference time of the forecast round (i.e., the time with the most recent available data point). The baseline assumes future changes follow the empirical distribution of recent 1-week differences. We define the set of recent weekly increments in the target quantity (e.g., ILI incidence) as: $\delta^{(c)} = \left\{y_s^{(c)} - y_{s-1}^{(c)} \;\; | \, s \;=\; t_0 + 1, \ldots, t_r\right\}$, where t_0 is the start of the calibration window (e.g., week 42 of 2023). To ensure that the median baseline forecast aligns with the last calibration point, we symmetrize δ by considering $\delta_{sym}^{(c)} = \delta^{(c)} \;\cup \left\{-d \mid d \;\in \delta^{(c)}\right\}$. From this symmetric set, we sample with replacement H increments $(d_1, \ldots, d_H)$ to construct predictive trajectories. The value of each trajectory at horizon h is computed as $\hat{y}_{t_r+h}^{(c)} = y_{t_r}^{(c)} + \sum_{i=1}^{h} d_i$, where $h \in \{1, \cdots, H\}$ and t_r is the reference time of the forecast round. We repeat this process to generate $N = 10{,}000$ synthetic trajectories, from which we extract forecast quantiles $F_{base,r,h}^{(c)}(q)$. Analogous baseline modelling approaches have been employed in other epidemiological forecasting challenges [5], [6], [15].

Forecast evaluation

To assess the performance of different forecasting models and the ensemble for different countries, forecasting horizons, and epidemic phases, forecasts are evaluated against reported surveillance data. We consider as *final* the data reported on week 16, 2024, which was the last week that RespiCast was open for submissions. We assess the predictive performance of models using the principle of producing the narrowest possible prediction intervals (i.e. maximizing the sharpness of predictive distributions) while maintaining proper calibration. This means that prediction intervals should be as narrow as possible and closely aligned with the real data, while ensuring that forecasted probabilities accurately reflect observed frequencies. The overall predictive performance is quantified using both the WIS and the AE of the median, while we assess calibration by means of the prediction coverage. Relative performance is expressed by means of the relative WIS and relative AE. In the next subsections, we briefly introduce these concepts.

Weighted Interval Score

The WIS is a scoring rule used to evaluate the accuracy of probabilistic forecasts [16]. It assesses the performance of quantile-based forecasts by penalizing both the distance between predicted and actual values, as well as under- or over-prediction. For a $(1 - \alpha) \times 100\%$ prediction interval of a forecast F, the interval score (IS) is defined as:

$$IS_{\alpha}(F, y) = \big(u(F) - l(F)\big) + \frac{2}{\alpha}(l(F) - y)\, I\big(y < l(F)\big) + \frac{2}{\alpha}\big(y - u(F)\big)\, I\big(y > u(F)\big),$$

where y is the observed outcome, $u(F)$ is the upper limit of the prediction interval corresponding to forecast F, $l(F)$ is the lower limit thereof, and $I(\cdot)$ is an indicator function that equals one when the specified condition is true, and zero otherwise. It follows that the first term measures the width of the prediction interval, while the second and third terms are penalties for, respectively, over- and under-prediction of the forecast weighted by the inverse of level α. The WIS is an extension of this metric to multiple intervals. More specifically, for K $(1 - \alpha_k) \times 100\%$ prediction intervals ($k = 1, \ldots, K$) of forecast F, the WIS is defined as:

$$WIS_{\alpha_{1:K}}(F, y) = \frac{1}{K + 1/2}\Big(w_0 |y - m| + \sum_{k=1}^{K} w_k IS_{\alpha_k}(F, y)\Big),$$

where m is the projected median and w_0, w_k are non-negative weights that we set, following a standard choice, to $1/2$ and $\alpha_k/2$, respectively. Here we consider $K = 11$ prediction intervals ($\alpha_k = 0.02, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90$), and WIS is computed for each forecast submission $F_{m,r,h}^{(c)}$ and the corresponding target outcome $y_{t_r+h}^{(c)}$, where r is the forecasting round, h is the forecast horizon, and c denotes the country.

Absolute Error

The AE is a commonly used metric to evaluate the accuracy of point forecasts. It measures the absolute difference between the predicted and the observed value, indicating how far off

a prediction is from the reported data. Given a forecast distribution F and observed value y, the AE is defined as:

$$AE(F, y) = |F(0.5) - y|,$$

where $F(0.5)$ denotes the median of the forecast distribution. Note that this metric solely measures the distance between the median projection and the observed value in absolute value in contrast with the WIS that also accounts for over- and under-prediction of the outcome. In our framework, AE is computed for each forecast submission $F_{m,r,h}^{(c)}$ and the corresponding target outcome $y_{t_r+h}^{(c)}$.

Prediction Coverage

Prediction coverage is a metric used to evaluate calibration of probabilistic forecasts, specifically by measuring the proportion of times the observed values fall within the predicted intervals (i.e. the nominal coverage). For a $(1 - \alpha) \times 100\%$ prediction interval, the prediction coverage over N forecast-observation pairs (F_i, y_i) is defined as:

$$Cov_\alpha = \frac{1}{N}\sum_{i=1}^{N} I(y_i \in [F_i(\alpha/2), F_i(1 - \alpha/2)]),$$

where y_i are the observed values, $F_i(\alpha/2)$ and $F_i(1 - \alpha/2)$ are the lower and upper bounds of the $(1 - \alpha) \times 100\%$ prediction interval for the i-th forecast, and $I(\cdot)$ equals one when the observed value falls within the prediction interval and zero otherwise.

A well-calibrated model will have prediction coverage scores closely matching the nominal coverage (e.g. the 90% of observed values fall in the 90th percentile prediction interval), while deviating values indicate over- or under-confidence. A model is overconfident if the prediction coverage is below the nominal coverage: in this case, the prediction interval is too narrow (e.g. only 80% of the observed values fall in the 90th percentile prediction interval). The reverse is true for an underconfident model: the prediction coverage is above the nominal coverage meaning the prediction interval is too wide.

Relative performance

To evaluate the performance of a model m relative to the baseline, we define the Relative Weighted Interval Score ($RelWIS$) as the base-2 logarithm of the ratio between the WIS of the baseline model and that of model m. For a given forecasting round r, horizon h, and country c, we compute:

$$RelWIS_{r,h}^{(c)}(m) = log_2\left(\frac{WIS\left(F_{base,r,h}^{(c)}, y_{t_r+h}^{(c)}\right)}{WIS\left(F_{m,r,h}^{(c)}, y_{t_r+h}^{(c)}\right)}\right).$$

Values larger than 0 indicate that the model performs better than the baseline, while values smaller than 0 indicate worse performance relative to the baseline. An analogous definition can be applied to the AE of the median. Unless stated differently, relative performance across different time horizons within the same forecasting round is averaged.

RespiCast forecasting platform

The RespiCast infrastructure consists of two primary components: a set of GitHub repositories with automated actions and a dedicated web platform focused on forecast visualization and evaluation. For each set of disease indicators, a corresponding repository is created and configured on GitHub. These repositories store ground truth data for model calibration and evaluation and support the submission and validation of forecasts from contributing teams. The repository structure adheres to the recommendations of the *hubverse* standard, a collection of open-source software and tools that support collaborative modeling challenges [17]. The RespiCast web platform features two interactive applications for exploring forecasts and evaluating model performance. The Forecasts app allows users to select targets, seasons, locations, and models via a menu on the left side, while a panel on the right allows for adjustments in visualization settings (e.g., scales, confidence intervals). The central chart compares surveillance data with selected forecasts, allowing for the comparison between models, including the RespiCast ensemble. The Evaluations app has a similar structure, enabling users to choose targets, metrics, and seasons. It provides two views: a summary table of average scores by model and location, and a detailed weekly evaluation view.

Figure 1. RespiCast forecasting platform [9].

Results

Participation and engagement

During the 2023/24 winter season, RespiCast collected forecasts for ILI and ARI incidence over 20 rounds in various European countries. As shown in Table 1, by the end of the season,

14 different forecasting models from 6 contributing teams were submitted to forecast ILI incidence, while results from 10 models from 3 teams were submitted for ARI incidence. On average, the outputs of 11 models were submitted per round for ILI incidence, and from 7 models for ARI incidence. The forecasts covered 27 European countries for ILI and 18 countries for ARI incidence. For a more detailed overview of the number of models over time, see Figure S15.

Forecasting Target	**Total number of Forecasting Models**	**Total number of Contributing Teams**	**Countries Covered**	**Average number of Models per Round**
ILI incidence	14	6	27	11
ARI incidence	10	3	18	7

Table 1: Summary of contributions to RespiCast for each forecasting target during 2023/2024 forecasting season.

In the following analysis, we focus on forecasts in countries where ground truth data was available for evaluation and at least three distinct models (excluding the ensemble and baseline) were submitted. This ensures that ensembles were generated from a minimum of three models. Additionally, forecasts for Latvia after 2024/04/21 were excluded due to anomalies in the reported data beyond that date, where incidences were consistently reported as 100,000 per 100,000. This leads to the inclusion of 26 countries for ILI incidence, and 14 for ARI incidence. It is important to note that not all countries reported surveillance data with the same frequency—some did not report data during certain weeks—and that England, Scotland, Northern Ireland, Wales, and Switzerland were only included in the set of countries accepting forecasts starting from forecasting round 11 (2024/02/28). Notably, no data was reported in week 52 due to the end-of-year holidays. In the Supplementary Information, we present the included ensemble forecasts for 1 to 4 weeks ahead (Figures S1 to S8).

In the Supplementary Information, we also provide a list of submitted models for each target, along with a brief description of their methodologies. The sample includes a diverse range of approaches, including statistical methods based on autoregressive models such as ARIMA, fully mechanistic models incorporating country-specific demographics and contact patterns, metapopulation models that account for individual mobility across and within countries alongside disease spread, and semi-mechanistic approaches based on effective reproductive number estimation, generalized logistic growth, and data integration techniques such as the extended Kalman filter. Additionally, some models employ weighted ensembles of individual approaches which are not among the models submitted independently by other participants.

Forecasting performance during the 2023/2024 winter season

Figure 2 illustrates the performance of ensemble forecasts for ILI incidence using the relative WIS (i.e., relative to the baseline model). Panel A presents the distribution of relative WIS values, aggregating results across all countries and forecasting rounds, separately for the ensemble and all other models. Interestingly, while individual models have a median relative WIS of 0.08, indicating marginal improvements over the baseline, the ensemble's median relative WIS is 0.45, demonstrating superior predictive performance for the ensemble as compared to the baseline and the individual models. The ensemble's distribution of relative WIS values is also narrower, with an interquartile range of 1.21 compared to 1.50 for the other models. This indicates that the ensemble is not only superior in terms of median performance but also exhibits greater stability and consistency across forecasting rounds and countries.

Performance appears to be heterogeneous in time and space. Panel B shows the temporal evolution of the relative WIS across epidemiological weeks. Here, the distribution of the relative WIS for a given week consists of relative WIS scores from forecasts produced in that week, averaged over the four-week forecasting horizon and aggregated across all countries. The figure reveals weaker ensemble performance during the early weeks of the RespiCast forecasting season, corresponding to the Christmas-New Year period. This may be attributed to several factors, including fewer model submissions, reduced data availability over the Christmas-New Year period and the noisy data at the early stage of the epidemic phase (see Discussion Section, Figure S15 and Figures S1-S4 for the ILI epidemic curves and corresponding ensemble forecasts). Despite these early challenges, the ensemble maintains a positive median relative WIS in all but six weeks, underscoring its overall reliability and robustness over time.

Panel C displays the distribution of the ensemble relative WIS across countries (aggregating scores from all weeks), ranked by median performance. The ensemble outperforms the baseline in over 90% of the countries, with a negative median relative WIS observed in only two (Estonia and Iceland). Notably, in 13 out of 26 countries the first quartile of the distribution is positive, meaning that the ensemble outperforms the baseline at least 75% of the times. Panel D provides a more granular view on relative WIS by country and week, showing the average relative WIS over the four-week forecasting horizon for forecasts produced in that week for each country. The picture emerging confirms distinct spatial and temporal performance patterns. Forecasts in some countries (e.g., Finland, Italy, Latvia) consistently show positive relative WIS across nearly all weeks, while others (e.g., Croatia, Czechia, Hungary) exhibit early challenges but improve over time. Conversely, in countries like Denmark, Estonia, and Iceland, the ensemble struggles to outperform the baseline. Across the board, the performance improves over time in all countries, with more blue-shaded cells appearing in later weeks (see Figures S1-S4 for the ILI epidemic curves and corresponding ensemble forecasts).

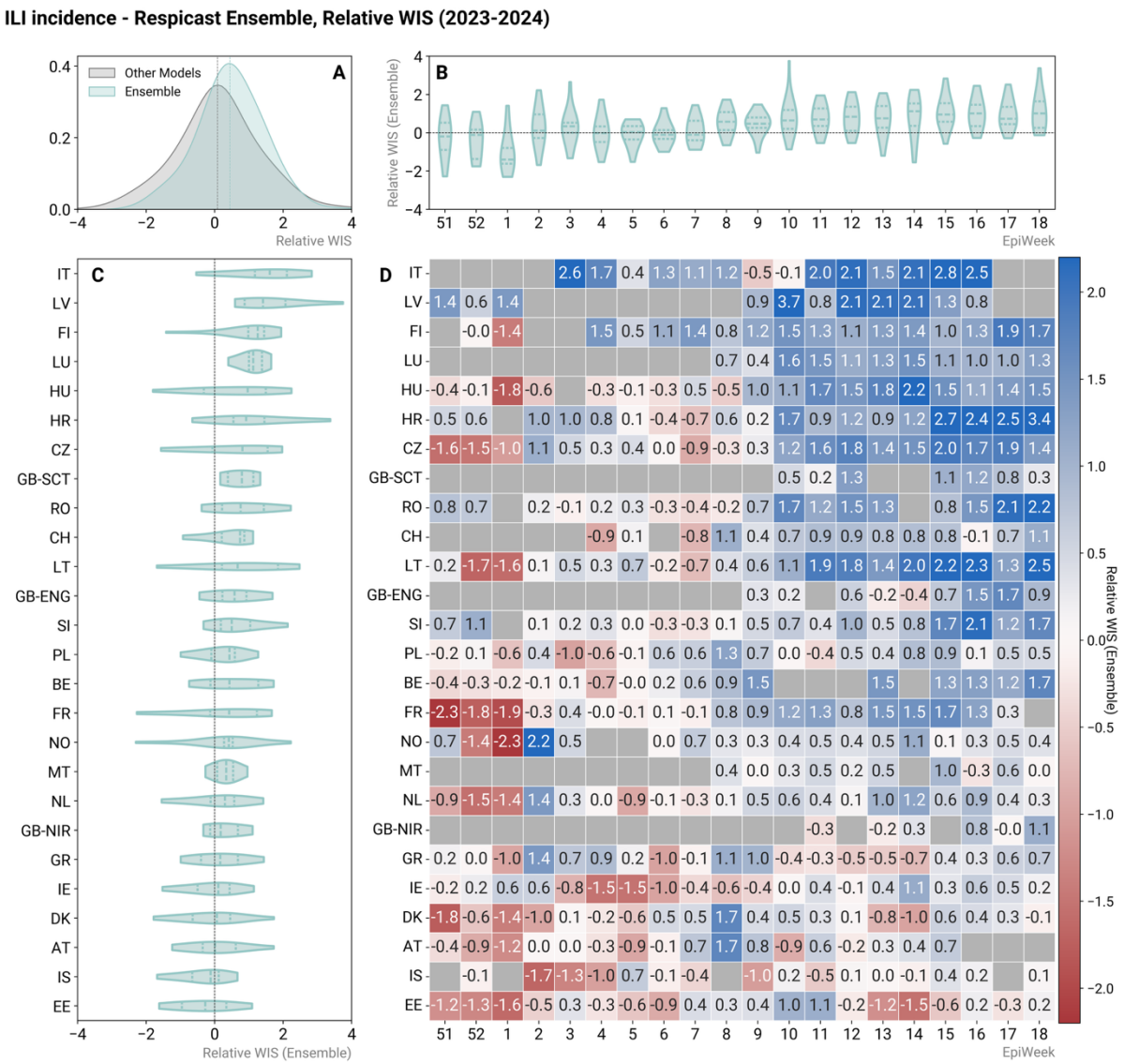

Figure 2. Relative Weighted Interval Score (WIS) for Influenza-Like Illness (ILI) incidence predictions from the RespiCast Ensemble during the 2023-2024 winter season. **(A)** Distribution of relative WIS across all countries and weeks, comparing the ensemble model to all other models combined. **(B)** Distribution of relative WIS of the ensemble across all countries for each epidemiological week separately. **(C)** Distribution of relative WIS of the ensemble across all epidemiological weeks for each country separately. **(D)** Heatmap of relative WIS values of the ensemble by country and epidemiological week, with blue indicating better performance (relative WIS greater than 0) and red indicating worse performance (relative WIS lower than 0) compared to the baseline. Grey cells indicate missing data. The violin plots in Panels B and C are truncated at the observed minimum and maximum values, with dashed lines inside representing the 1st quartile, median, and 3rd quartile of the distribution. In all panels, relative WIS values are averaged over the four-week forecasting horizon.

In Figure 3, we replicate the analysis for ARI incidence, revealing overall consistency with the findings for ILI. The ensemble again demonstrates performance improvements in terms of relative WIS compared to individual models (Figure 3A). However, its median relative WIS is 0.27, lower than the 0.45 observed for ILI incidence. Performance patterns across locations and times align with those seen for ILI incidence. Ensemble forecasts produced in earlier weeks exhibit lower relative WIS, with performance improving as the season progresses (Figure 3B). Across countries, only Finland has a negative median relative WIS (Figure 3C).

This indicates that the ensemble outperforms, in median terms, the baseline in all other locations. Finally, Figure 3D highlights the persistent heterogeneity in patterns across locations and times, with a clear trend of improving performance over the weeks and certain countries consistently facing greater forecasting challenges, such as Luxembourg, Romania, and Finland.

Overall, these results highlight the ensemble's predictive improvements, while also identifying specific regions and periods where its accuracy remains challenging. In the Supplementary Information, we replicate the analyses presented in Figures 2 and 3, using relative AE instead of relative WIS as scoring rule. Overall, the results remain consistent, confirming the ensemble's improved forecasting performance compared to the baseline. However, the magnitude of improvement is more modest when evaluated using the relative AE rather than relative WIS. Additionally, in the Supplementary Information, we compare the ensemble to individual models using both relative WIS and AE across the two forecasting targets. The results show that the ensemble generally outperforms individual models, with the performance advantage being more pronounced when using relative WIS as the scoring rule.

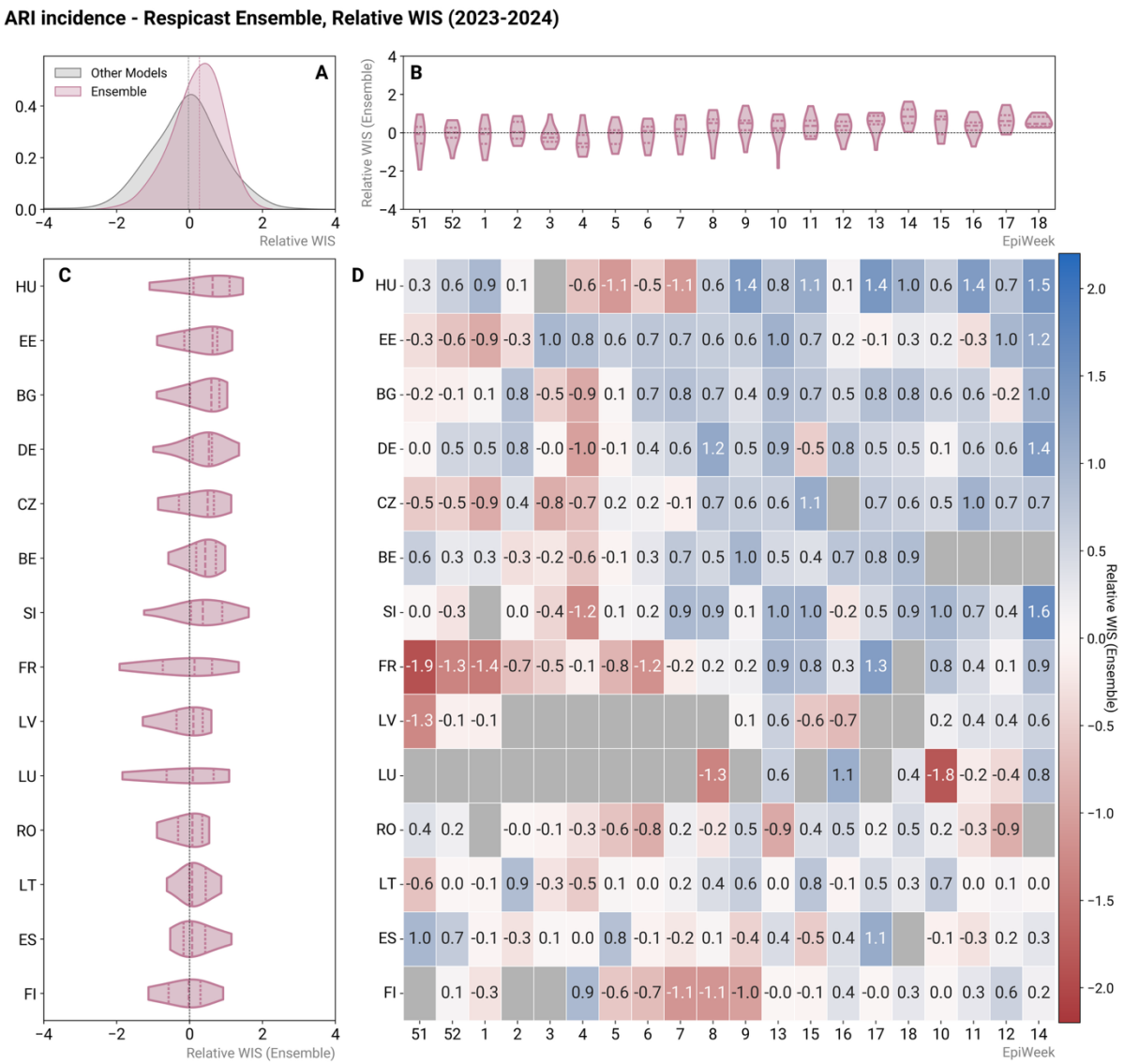

Figure 3. Relative Weighted Interval Score (WIS) for Acute Respiratory Infection (ARI) incidence predictions from the RespiCast Ensemble during the 2023-2024 winter season. **(A)** Distribution of relative WIS across all countries and weeks, comparing the ensemble model to all other models combined. **(B)** Distribution of relative WIS of the ensemble across all countries for each epidemiological week separately. **(C)** Distribution of relative WIS of the ensemble across all epidemiological weeks for each country separately. **(D)** Heatmap of relative WIS values of the ensemble by country and epidemiological week, with blue indicating better performance (relative WIS greater than 0) and red indicating worse performance (relative WIS smaller than 0) compared to the baseline. Grey cells indicate missing data. The violin plots in Panels B and C are truncated at the observed minimum and maximum values, with dashed lines inside representing the 1st quartile, median, and 3rd quartile of the distribution. In all panels, relative WIS values are averaged over the four-week forecasting horizon.

In Figure 4 we evaluate the calibration of forecasts by comparing prediction coverage (y-axis) with nominal coverage (x-axis) across the four forecasting horizons separately (1 to 4 weeks ahead) for each forecasting target. Panels A and B show results for ILI and ARI incidence, respectively. Each point represents the prediction coverage obtained pooling together forecasts across all countries, with the dashed diagonal line indicating perfect calibration (i.e., where prediction intervals match nominal coverage). It follows that points above the diagonal indicate underconfident forecasts, where prediction intervals are too wide, and prediction

coverage exceeds the nominal level. Conversely, points below the diagonal suggest overconfident forecasts, where prediction intervals are too narrow, leading to underestimation of uncertainty and lower-than-expected prediction coverage. The extent of deviation from the diagonal reflects the magnitude of miscalibration.

For ILI incidence, the ensemble forecasts align well with the diagonal across horizons at lower nominal coverage levels (i.e., up to 50%), indicating good overall calibration in that range. For higher nominal coverage levels, a consistent trend of overconfidence is observed, with the magnitude of undercoverage increasing with horizon length. For example, at 90% nominal coverage the prediction coverage is 79% at horizon 1 (-11% deviation from nominal coverage), which reduces to 75% at horizon 4 (-15% deviation). Nonetheless, the ensemble shows improved calibration across all levels and horizons compared to the pooled set of all other models. ARI incidence exhibits a comparable pattern, with the ensemble forecasts showing improved calibration relative to individual models. However, systematic undercoverage remains evident across all forecast horizons. This undercoverage is generally more consistent at higher nominal coverage levels. Moreover, no systematic degradation of calibration is observed as the forecast horizon increases. By contrast, the baseline model shows relatively good calibration in both cases and even outperforms the ensemble in certain instances, though it also displays signs of underconfidence, particularly when predicting ARI incidence. While the pooled analysis indicates overall reliable calibration for the ensemble, country-level results reveal heterogeneity. Supplementary Figure S14 summarizes deviations in 90% prediction coverage by country, highlighting substantial location-specific variability in ensemble forecast calibration.

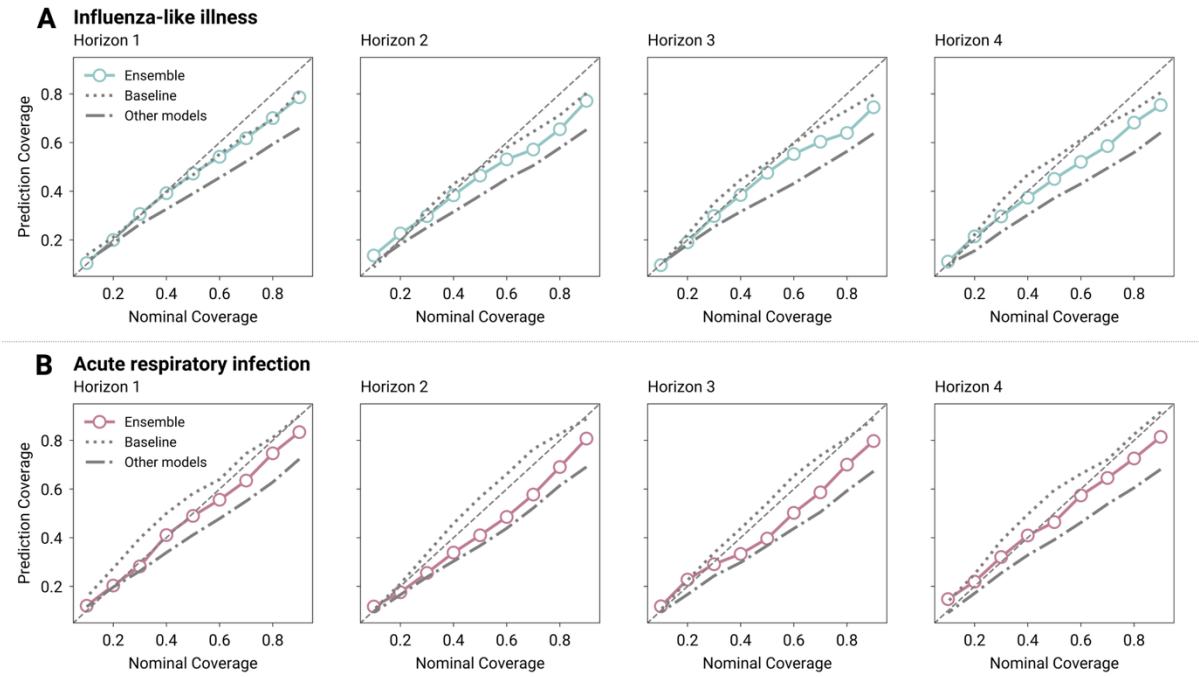

Figure 4. A) Nominal versus ensemble prediction coverage across forecasting horizons from 1 to 4 weeks ahead, for ILI (A) and ARI (B) incidence forecasting targets. Results are shown for the ensemble (colored circles), baseline (dotted line), and single models (dash-dot line). Ideal calibration corresponds to the diagonal line.

Discussion

We evaluated the performance of ensemble forecasts in the context of the new ECDC forecasting hub, RespiCast, during the 2023/24 winter season. The analysis shows the effectiveness of the ensemble in improving forecast accuracy and reliability compared to a naive baseline model for ILI and ARI respiratory disease indicators across Europe. The ensemble outperformed the baseline model at most times and for most countries, and demonstrated greater reliability compared to individual models. The performance analysis also revealed challenges, mainly variation in forecasting performance over time and by country as well as overconfidence in the ensemble forecasts.

Forecasting performance (measured as relative WIS and AE compared to a naive baseline model) varied over time, with lower scores early in the forecasting season (December-January), followed by improvements as the season progressed. Several factors may have contributed to the initial lower scores. First, this period coincided with an increase or peak in ILI and/or ARI incidence in most countries, which represents one of the most uncertain phases of the epidemic to model. Second, fewer models were submitted during the initial rounds (as shown in Figure S15), potentially reducing the ensemble's performance, as ensemble performance is closely tied to the number of contributing models [18]. Future analyses could examine how model participation and methodological heterogeneity influence RespiCast ensemble performance across epidemic phases. In this context, a key open question is how to optimize ensemble composition. Understanding the role of model type, quality, and complementarity in shaping ensemble effectiveness could inform strategies for model

selection and weighting. Actively managing contributions - by promoting participation during critical periods and encouraging methodological diversity - may further improve ensemble performance, ultimately increasing the reliability and utility of forecasts for public health decision-making. Third, the start of the RespiCast season coincided with the Christmas period, a time when reporting may be delayed and less accurate, potentially affecting model calibration and forecasting performance. Another factor to consider is that forecasters may have improved their models over time, as they gained experience with the forecasting system and the evolving epidemic context [19]. It is also important to note that the 2023/24 winter season marked RespiCast's inaugural year, and its initial performance may have been affected by a learning curve and limited data availability for early-season model calibration. Since these factors overlapped in the first season, further multi-season evaluation will be needed to disentangle their individual contributions to ensemble performance. It is important to acknowledge that the performance metrics used in this analysis are defined relative to the baseline model, and their interpretation may be influenced by the baseline's behaviour as well as the characteristics of the underlying data. For instance, performance during periods of rapid change or high incidence may affect both the baseline and ensemble scores in non-trivial ways.

RespiCast provided forecasts for 27 European countries, revealing substantial variability in performance across locations. This indicates the potential impact of differing local epidemic dynamics or differences in surveillance systems on forecast accuracy. Identifying the specific country-level factors driving these differences remains challenging without extensive local knowledge and multi-season data. Additionally, the option for modeling teams to selectively submit forecasts across countries and rounds may introduce bias, as models might opt out of settings where they previously underperformed, potentially skewing ensemble reliability in some locations. In the future, with additional seasons of forecasts being available, we may be able to further investigate the underlying drivers of this heterogeneous performance. Forecast accuracy is also inherently tied to the availability and quality of underlying data. Differences in data completeness, timeliness, and stability across countries can significantly affect model calibration and evaluation. For instance, irregular updates or retrospective revisions in surveillance data may distort accuracy of forecasts or limit their real-time utility. These challenges underscore the need for continued investment in harmonized and high-quality surveillance systems across Europe.

Furthermore, ensemble forecasts exhibit overconfidence, particularly at longer prediction horizons, likely due to overconfidence in individual models. Addressing this issue could involve requesting teams with systematically overconfident models to reassess their calibration or developing new ensembling methods that adjust for the historical accuracy of individual models' forecast intervals to improve reliability.

Another consideration relates to the choice of forecasting targets. While ILI and ARI are widely reported syndromic indicators that enable consistent pan-European forecasting, they also present limitations when it comes to forecasting as neither are pathogen-specific. Forecasting pathogen-specific targets may be conceptually clearer and more directly linked to transmission dynamics. For example, syndromic indicators could be multiplied by pathogen-specific

positivity rates to serve as pathogen-specific incidence proxies in the future, similar to the Goldstein Index frequently used as a proxy for influenza activity [22]. This pathogen-specific approach could also support the evolution of RespiCast into a more integrated forecasting hub aligned with the multi-pathogen surveillance data reported within ERVISS, enhancing its value for situational awareness and public health decision-making.

Further work is needed to align forecasting efforts with public health needs. This includes, for example, refining evaluation metrics to be more relevant for decision-makers and improving forecasting performance during critical epidemic phases. A user-focused assessment could guide next steps and help tailor future iterations of RespiCast to better support public health surveillance and response.

In conclusion, the RespiCast ensemble model outperformed individual models and the baseline model during 2023/24 winter season in terms of WIS and absolute error. This season served as a valuable proof of concept, identifying key areas for improvement, such as accuracy at longer horizons, calibration, and variability across times and locations. RespiCast establishes a solid foundation for advancing forecasting capabilities in Europe. Building on the lessons learned, RespiCast has continued into the 2024/25 season with increased participation from modelling teams and additional forecasting targets are being explored to further develop its potential as a tool for supporting public health decision-making.

Data Availability

All the data used in this manuscript are available online at:

- ILI: https://github.com/european-modelling-hubs/flu-forecast-hub_archive
- ARI: https://github.com/european-modelling-hubs/ari-forecast-hub_archive

Acknowledgements and Funding

Data acknowledgment:

- For EU/EEA countries - ERVI-Net (European Respiratory Virus Surveillance Network) reporting data to TESSy.
- For UK & Switzerland: fluID (WHO)

This work received funding from the European Centre for Disease Prevention and Control (ECDC) via contract ECDC/2023/045 LOT1. S.FU. and K.S. are supported by Wellcome Trust (210758/Z/18/Z). A.A. is supported by the Luxembourg Government through the CoVaLux programme (16954531). D.PR. and G.G. are funded by the European Union through the ERC INSPIRE grant (project number 101076926). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the European Research Council Executive Agency can be held responsible for them. Y.L. acknowledges support from the China Scholarship Council (CSC). Y.L. and N.P. thank the High-Performance Computing facilities at Queen Mary University of London. All authors thank Dr. Johannes Bracher for the useful discussion.

References

[1] Desai AN, Kraemer MU, Bhatia S, Cori A, Nouvellet P, Herringer M, Cohn EL, Carrion M, Brownstein JS, Madoff LC, Lassmann B. Real-time epidemic forecasting: challenges and opportunities. Health security. 2019 Aug 1;17(4):268-75.

[2] Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, Fernandez D, Greene SK, Kipshidze N, Liu L, Mgbere O. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health. 2019 Dec 10;19(1):1659.

[3] Reich NG, Lessler J, Funk S, Viboud C, Vespignani A, Tibshirani RJ, Shea K, Schienle M, Runge MC, Rosenfeld R, Ray EL. Collaborative hubs: making the most of predictive epidemic modeling. American Journal of Public Health. 2022 Jun;112(6):839-42.

[4] Shea K, Runge MC, Pannell D, Probert WJ, Li SL, Tildesley M, Ferrari M. Harnessing multiple models for outbreak management. Science. 2020 May 8;368(6491):577-9.

[5] Sherratt K, Gruson H, Johnson H, Niehus R, Prasse B, Sandmann F, Deuschel J, Wolffram D, Abbott S, Ullrich A, Gibson G. Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. Elife. 2023 Apr 21;12:e81916.

[6] Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, Gerding A, Gneiting T, House KH, Huang Y, Jayawardena D. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences. 2022 Apr 12;119(15):e2113561119.

[7] Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, Osthus D, Ray EL, Tushar A, Yamana TK, Biggerstaff M. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proceedings of the National Academy of Sciences. 2019 Feb 19;116(8):3146-54.

[8] European respiratory virus surveillance summary (ERVISS). [Online]. Available: https://erviss.org/

[9] RespiCast website. [Online]. Available: https://RespiCast.ecdc.europa.eu/

[10] RespiCast GitHub repository (ILI incidence). [Online]. Available: https://github.com/european-modelling-hubs/flu-forecast-hub_archive

[11] RespiCast GitHub repository (ARI incidence). [Online]. Available: https://github.com/european-modelling-hubs/ari-forecast-hub_archive

[12] Operational considerations for respiratory virus surveillance in Europe. Copenhagen: WHO Regional Office for Europe and Stockholm: European Centre for Disease Prevention and Control; 2022.

[13] 'WHO global influenza programme FluID'. [Online]. Available: https://www.who.int/teams/global-influenza-programme/surveillance-and-monitoring/fluid

[14] Howerton E, Runge MC, Bogich TL, Borchering RK, Inamine H, Lessler J, Mullany LC, Probert WJ, Smith CP, Truelove S, Viboud C. Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology. Journal of the Royal Society Interface. 2023 Jan 25;20(198):20220659.

[15] Fiandrino S, Bizzotto A, Guzzetta G, Merler S, Baldo F, Valdano E, Mateo Urdiales A, Bella A, Celino F, Zino L, Rizzo A, Li Y, Perra N, Gioannini C, Milano P, Paolotti D, Quaggiotto M, Rossi L, Vismara I, Vespignani A, Gozzi N. Collaborative forecasting of influenza-like illness in Italy: The Influcast experience. Epidemics. 2025;50:100819. doi:10.1016/j.epidem.2025.100819.

[16] Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLoS Comput Biol. 2021;17(2):e1008618.

[17] 'The hubverse: open tools for collaborative modeling'. [Online]. Available: https://hubverse.io/en/latest/

[18] Fox SJ, Kim M, Meyers LA, Reich NG, Ray EL. Optimizing disease outbreak forecast ensembles. Emerg Infect Dis. 2024;30(9):1967.

[19] McGowan CJ, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Sci Rep. 2019;9(1):683.
[20] Yardley L, Petropoulos F. Beyond error measures to the utility and cost of the forecasts. Foresight Int J Appl Forecast. 2021;63:36–45.
[21] Gerding A, Reich NG, Rogers B, Ray EL. Evaluating infectious disease forecasts with allocation scoring rules. J R Stat Soc A. 2024;qnae136.
[22] Goldstein E, Cobey S, Takahashi S, Miller JC, Lipsitch M. Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: a statistical method. PLoS medicine. 2011 Jul 5;8(7):e1001051.