ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Integrating nowcasts into an ensemble of data-driven forecasting models for SARI hospitalizations in Germany

Daniel Wolffram [a,b], Johannes Bracher [a,b], RESPINOW Study Group [1], Melanie Schienle [a,b,*]

[a] Karlsruhe Institute of Technology, Germany
[b] Heidelberg Institute for Theoretical Studies, Germany

ARTICLE INFO

Keywords:
Integration of probabilistic nowcasts and forecasts
Short-term now- and forecasting
Multi-model approach
Respiratory diseases
Hospitalization incidence

ABSTRACT

Predictive epidemic modeling can enhance situational awareness during emerging and seasonal outbreaks and has received increasing interest in recent years. A common distinction is between nowcasting, which corrects recent incidence data for reporting delays, and forecasting, which predicts future trends. This paper presents an integrated system for nowcasting and short-term forecasting of hospitalizations from severe acute respiratory infections (SARI) in Germany (November 2023–September 2024). Motivated by facilitating multi-model forecasting collaborations, we propose a modular approach in which a statistical nowcasting model is run centrally, and its output is provided as input to various data-driven forecasting methods. We apply this approach to a seasonal time series model, a gradient boosting approach, and a neural network. These are moreover combined into an ensemble approach, which achieves the best average performance. The resulting forecasts are overall well-calibrated up to four weeks ahead, but struggled to capture the unusual double peak that occurred during the test season. The presented retrospective results are key developments for ongoing and future collaborative real-time forecasting of respiratory diseases in Germany.

1. Introduction

Predictive modeling of infectious diseases has received considerable attention in recent years, fueled by the public health crises of COVID-19 (e.g., Bracher et al. 2021, Cramer et al. 2022) and mpox (e.g., Bleichrodt et al. 2024). Disease forecasting is a broad field, and three main types of predictive modeling tasks can be distinguished (Reich et al. 2022, see Fig. 1).

- **Nowcasting** is the statistical correction of recent data points that are yet incomplete and subject to delayed additions. Nowcasts, hence, refer to recent rather than upcoming infection dynamics, but are predictive in that they anticipate data revisions and reveal current trends.
- **Short-term forecasts** are unconditional predictions about the future course of an epidemic. These are

* Corresponding author at: Karlsruhe Institute of Technology, Germany.
E-mail address: melanie.schienle@kit.edu (M. Schienle).

[1] The RESPINOW Study Group are: Alex Dulovic, Alexander Kuhlmann, André Karch, Berit Lange, Carolina Klett-Tammen, Chao Xu, Claudia Denkinger, Cornelia Gottschick, Daniel Wolffram, Felix Guenther, Florian Marx, Isti Rodiah, Johannes Bracher, Laura-Inés Boehler, Manuela Harries, Melanie Schienle, Michael Böhm, Nicole Schneiderhan-Marra, Nils Bardeck, Olga Hovardovska, Patrick Marsall, Philipp Dönges, Rafael Mikolajczyk, Rolf Kaiser, Sebastian Contreras, Torben Heinsohn, Tyll Krüger, Ulrich Reinacher, Veronika K. Jaeger, Viola Priesemann, Wolfgang Bock.

The numerical results presented in this manuscript were reproduced by the Editor-in-Chief on 10 January 2026.

https://doi.org/10.1016/j.ijforecast.2026.01.001

Please cite this article as: D. Wolffram, J. Bracher, M. Schienle, Integrating nowcasts into an ensemble of data-driven forecasting models for SARI hospitalizations in Germany. International Journal of Forecasting (2026), https://doi.org/10.1016/j.ijforecast.2026.01.001.

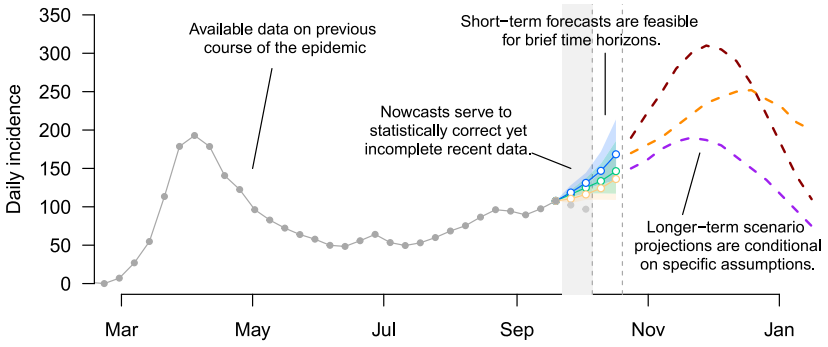

Fig. 1. Distinguishing nowcasting, short-term forecasting, and scenario modeling of infectious diseases.

feasible only for short time periods, with appropriate prediction horizons depending on the type of indicator to predict.

- **Scenario projections** are used to make statements about possible longer-term developments, but are conditional on explicit assumptions that may or may not correspond to the future conditions encountered in the real world. For instance, scenarios may elucidate possible epidemic trajectories under various intervention strategies.

While scenario modeling has a somewhat different focus and purpose, nowcasting and short-term forecasting boil down to the same task: generating probabilistic statements about disease incidence at various points in time. For all three tasks, multi-model approaches have been found particularly suitable (Reich et al., 2022). The presence of multiple distinct models enables more realistic assessments of the predictive uncertainty and can be the basis for ensemble forecasts, which have often been found to be more robust (see e.g., Cramer et al. 2022).

In the United States, multi-model forecasting in collaborative *Forecast Hubs* was established in the early 2010s, most prominently for seasonal influenza (Reich et al., 2019). In many other countries, such systems were first implemented during the COVID-19 pandemic (Funk et al. 2021, Bracher et al. 2021, Paireau et al. 2022, Sherratt et al. 2023). To preserve the capacities built during the pandemic, these efforts now face the challenge of transitioning to routine operations in a seasonal rather than an emerging disease setting (see, e.g., Fi et al. 2025). This raises multiple challenges, but also offers the opportunity to revise and refine previous approaches. The present paper proposes a multi-model prediction system for severe acute respiratory infections (SARI) in Germany, which is the backbone of a new operational forecasting platform (see Section 5). Guided by our application setting, we focus on three aspects that, in combination, constitute the novelty of our contribution.

Firstly, post-COVID-19 forecasting efforts need to adapt to different, coarser data streams (Mathis et al., 2024), often with less timely reporting. Unlike predecessors, which focused either on nowcasting (Wolffram et al., 2023) or forecasting (Bracher et al., 2021), the new platform integrates both tasks, a novelty for collaborative efforts. Previous Forecast Hubs circumvented the need for nowcasting by aggregating incidence counts according to the date of report rather than, e.g., symptom onset. This, however, blurs recent trends and is poorly motivated from an epidemiological standpoint. Alternatively, the most recent data points can be removed entirely (see e.g., Paireau et al. 2022), but this implies discarding valuable information.

Secondly, while the COVID-19-related efforts in Germany were based on symptom-specific indicators in an acute outbreak setting, we deal with syndromic indicators in a seasonal context. Rather than mechanistic compartmental models, we therefore explore a range of phenomenological approaches to capture short-term dependencies and seasonal variation (see e.g., Albrecht et al. 2024, Brooks et al. 2018 for related work). We cover conceptually diverse modelling paradigms by considering a seasonal count time series model (Bracher & Held, 2022), a gradient boosting approach (Ke et al., 2017), and a neural network (Chen et al., 2023). These are moreover combined into an ensemble and compared to flat-line and seasonal benchmark models.

Lastly, we face the difficulty that the COVID-19 pandemic not only increased the overall respiratory disease burden but also altered the dynamics of other respiratory diseases (see, e.g., Buchholz et al. 2023). This is true for the years 2020–2022, when the associated non-pharmaceutical interventions largely stopped the spread of other respiratory diseases, as well as for the following period, when the immunity landscape was considerably different from earlier years. We will compare different approaches to using historical data from these periods for model fitting.

All challenges are addressed with collaborative multi-model forecasting in mind. As nowcasting and the necessary handling of multiple data versions impose significant overhead on participating forecasting teams, we develop a modular system in which the nowcasting task is split off, and nowcasts from a simple nowcasting model are provided via the Hub infrastructure. These can then be fed into diverse forecasting models. Moreover, this modular approach enables detailed diagnostics of how different approaches to handling reporting delays affect predictive performance. While the current evaluation is retrospective, the platform has transitioned to real-time operations in Fall 2024, and a prospective evaluation study for multiple surveillance indicators has been preregistered (Bracher & Wolffram, 2024).

Related works in the literature include the pre-COVID-19 works by Brooks et al. (2018) and Osthus et al. (2019), who combine nowcasts based on auxiliary data streams (rather than partial observations) with prediction models. Approaches to handle delayed reporting have been discussed by Ma et al. (2024) and De Nicola et al. (2022), who evaluate point predictions, and by Beesley et al. (2022) and Charniga et al. (2024), who consider probabilistic forecasts.

We find all our forecasting models to be well-calibrated for total weekly hospitalization incidences (coverage of 95% prediction intervals mostly between 90% and 95%). In the age-stratified setting, only the ensemble achieves nominal coverage, while some individual models drop to around 80% coverage. We note, however, that these coverage levels are achieved with relatively wide uncertainty intervals surrounding peak weeks, and there are difficulties in dealing with the double peak occurring in the test season. All three models considered outperform simple baseline models, though the differences are not consistently statistically significant across lead times. Similar to previous studies, the ensemble approach achieves the best overall performance in terms of the weighted interval score. Including a nowcasting step improves forecasts relative to a procedure that simply discards the most recent data point. Indeed, the loss in forecasting performance relative to a hypothetical setting in which the data are not subject to reporting delays is minor. This leads us to recommend incorporating nowcasting steps into infectious disease forecasting systems.

The remainder of the paper is structured as follows. Section 2 provides background information on SARI hospitalizations in Germany. In Section 3, we define the nowcasting and forecasting targets and present the methods employed for both tasks. Particular attention will be paid to how to feed nowcast information into forecasting models while accounting for the uncertainties that arise. In Section 4, we evaluate the resulting probabilistic forecasts visually and with a variety of metrics. Section 5 concludes with a discussion and a brief outlook. All results in this paper can be reproduced using the publicly available replication package at https://github.com/dwolffram/replication-sari-forecasting.

2. The SARI hospitalization incidence

2.1. Definition and description

Respiratory disease activity in Germany is monitored by a multitude of surveillance systems, including mandatory reporting schemes and virological and syndromic surveillance (Goerlitz et al., 2021). In the present paper, we focus on the incidence of hospitalization for *severe acute respiratory infections* (SARI). Since fall 2014, data on such hospitalizations have been collected in the *ICOSARI* system operated by the Robert Koch Institute (RKI; Buda et al. 2017, Tolksdorf et al. 2022). They are publicly accessible via the RKI GitHub repository (https://github.com/robert-koch-institut/SARI-Hospitalisierungsinzidenz). The SARI hospitalization incidence is a *syndromic indicator*, i.e., the case definition is based on the symptoms patients present rather than laboratory testing for a given pathogen. Specifically, a set of ICD-10 diagnostic codes (J09–J22) is used, see Buda et al. (2017) for details. Data collection is carried out via a sentinel system comprising roughly 70 hospitals across 13 of the 16 German federal states. The system covers around 6% of all hospitalizations occurring in Germany. Based on information on the catchment population covered by the sentinel sites, the SARI hospitalization incidence per 100,000 inhabitants can be estimated. Estimates at a weekly resolution (with weeks starting on Mondays) are available both unstratified (00+) and by six age groups (0–4, 5–14, 15–34, 35–59, 60–79, 80+). In this paper, we rescale the estimated incidence to absolute count values.

The pooled and age-group-wise incidence time series for the period 2014–2024 are displayed in Fig. 2 (see Supplementary Figures S2 and S3 for descriptive plots of the autocorrelation functions). Seasons we consider substantially affected by the acute phase of the COVID-19 pandemic are delimited by dashed vertical lines. Colors indicate the split into training, validation, and test periods, see Section 4.2.4 for details. Especially, in the age groups 05–14, 15–34, and 35–59, the test season displays rather unusual patterns, with consistently high incidences even in late spring and summer. In the very young and the elderly, this is less pronounced. Because these age groups have higher absolute numbers, the pooled incidence shown in the left panel exhibits a more typical seasonal pattern.

2.2. Data revisions and reporting delays

Like many epidemiological indicators, the SARI hospitalization incidence is subject to retrospective data revisions. Typically, the numbers are corrected upwards as additional hospitalizations are reported with a delay. To assess the impact of reporting delays, archives of historical data snapshots are necessary. The public RKI GitHub repository contains such snapshots back to the data release on 28 September 2023. Before this date, PDF reports were made available, which enabled the reconstruction of snapshots at the aggregate level back to early 2023 (though not for the different age groups).

As illustrated in the left panel of Fig. 3, reporting delays lead to an artificial dip at the end of the real-time incidence time series. Once data points have been completed over the following weeks, this dip disappears, and the actual trend becomes visible. For the SARI data, corrections become largely negligible after three weeks. The right panel shows the completeness of the data, zero to four weeks after the initial release, by data release week. It can be seen that, on average, initial data releases contain roughly 75% of the hospitalizations (or, put differently, initial values are corrected upwards by roughly a third). Initial reporting completeness fluctuates somewhat over time. Between Christmas and New Year, no releases occur, so all hospitalizations from this period are reported with a delay.

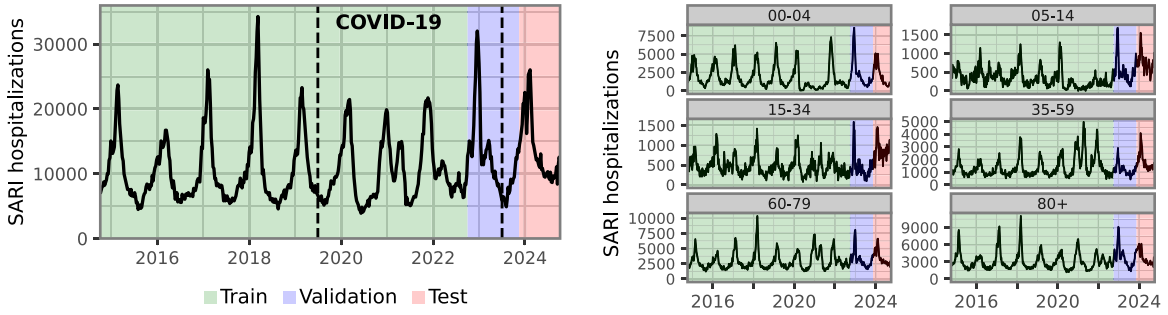

Fig. 2. Time series of weekly SARI hospitalizations in Germany, 2014–2024. Colors indicate the split of the data into training, validation, and test data; see details in Section 3.4.2. The portion labeled "COVID-19" is only included in the training set for part of our model specifications. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

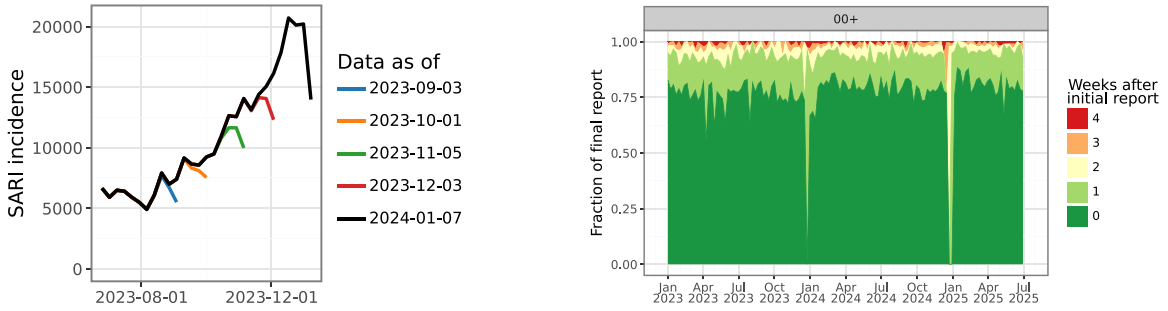

Fig. 3. Left: Illustration of data revisions in the SARI hospitalization incidence. Time series available on different dates are shown in different colors, overlaid with more complete data in black. A continued upward trend in the revised data replaces the apparent downward trend in the initial data versions. Right: Completeness of SARI hospitalization data zero to four weeks after the first release, per week (2023–2024). In alignment with the nowcasting target definitions (see Section 3), we consider only delays up to four weeks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.3. Auxiliary data

Some of the considered forecasting models use an auxiliary data set on weekly outpatient consultations for acute respiratory infections (ARI). These are based on a separate sentinel network of general practitioners, and data collection is likewise coordinated by the Robert Koch Institute (Goerlitz et al., 2021). The ARI time series shows seasonal patterns similar to those of SARI. Visual inspection reveals that it sometimes leads by a short time difference, which may make it a helpful predictor. More details on this data set are provided in Supplement B, and a visualization is shown in Figure S1.

3. Methods

3.1. Definition of the nowcasting and forecasting tasks

Nowcasting addresses the statistical correction of reporting delays as described for the SARI data in Section 2.2. Forecasting concerns the future epidemiological development and thus time points for which not even partial data is currently available.

We now generate weekly nowcasts and forecasts for the period from 16 November 2023 through 12 September 2024, following the data release schedule on Thursdays. We skipped Thursday, 28 December 2024, as there was no data release available. This *test period* is highlighted in red in Fig. 2. Counting from the day of data release (Thursday), the week ending on the preceding Sunday is indexed as *horizon* or *lead time* 0 weeks. Nowcasts, i.e., corrections of available preliminary data for, e.g., reporting delays, are produced for weeks −3 through 0. Forecasts are generated for horizons 1 through 4. All predictions are generated for the total weekly number of SARI hospitalizations at the national level (aggregated across all ages) and stratified by age group. We note that the available SARI hospitalization incidence is an estimate (see the previous section). In practice, we neglect any uncertainty associated with these estimates and treat them as the observable prediction target.

In the presence of data revisions, the definition of the prediction targets requires specific care. Based on experience from previous work (Wolffram et al., 2023), we define the final data version against which both nowcasts and forecasts are evaluated via a maximum reporting delay of $D = 4$ weeks. For each week, the respective data point used in the evaluation is thus set to the value available after four weeks of revisions (i.e., as published four weeks after the first data release containing a value for the respective week). This definition has the advantage of providing a well-defined target, with all observations in the evaluation period given the same amount of time for revisions. It is, however, unusual in that the time series

used for evaluation is not identical to any specific public data release.

For each nowcast or forecast horizon, we collect predictive quantiles at levels 2.5%, 10%, 25%, 50%, 75%, 90%, and 97.5%. This storage format corresponds to that of various Forecast Hubs established during the COVID-19 pandemic (Cramer et al., 2022; Wolffram et al., 2023).

3.2. Evaluation metrics

The primary evaluation metric is the *weighted interval score* (WIS, Bracher et al. 2021), which can be expressed as a sum of pinball losses. For quantiles $q_1, q_2, \dots q_K$ at levels $\tau_1 < \tau_2 < \dots < \tau_K \in (0, 1)$ and an observed value y it is given by

$$\text{WIS}(q_1, \dots, q_K; y) = \frac{1}{K} \sum_{k=1}^{K} 2 \times (\mathbf{1}\{y < q_k\} - \tau_k) \times (q_k - y),$$

where $\mathbf{1}$ denotes the indicator function. In our application, we use the previously mentioned levels 2.5%, 10%, 25%, 50%, 75%, 90%, 97.5%. We note that an alternative definition via so-called interval scores exists (hence the name; see Bracher et al. 2021). This display allows for a decomposition into components for forecast dispersion, overprediction, and underprediction, which we will use to enhance the interpretability of performance summary plots.

The WIS is negatively oriented, meaning that lower values are better. It can be seen as a probabilistic extension of the absolute error and approximates the commonly used continuous ranked probability score (CRPS, Gneiting et al. 2005). It is a proper scoring rule, thus incentivizing honest forecasting. Significance of score differences is assessed using Diebold–Mariano tests (Diebold & Mariano, 2002).

As a secondary performance metric, we use absolute errors of predictive medians to assess the quality of point forecasts. To assess forecast calibration separately, the empirical coverage proportions of predictive 50% and 95% prediction intervals are reported. These are given by the fraction of instances in which a prediction interval with a given nominal coverage contained the observed value.

Lastly, we complement our evaluation with an application of the recently proposed Rank Graduation Accuracy measure (RGA; Giudici and Raffinetti 2025). RGA is based on comparisons between the ranks of predicted and observed values and generalizes the commonly used area under the Receiver Operating Characteristic (ROC) curve to quantitative prediction targets. Intuitively, the RGA summarizes the concordance between the rank structure of the observations and point predictions. RGA values are contained in the unit interval, with a value of 1 indicating perfect rank concordance. Details on this metric are provided in Supplement D.

3.3. Nowcasting method and the coupling of nowcasting and forecasting

We separate the nowcasting and forecasting steps and use a separate nowcasting model that provides input to several forecasting models. While it may seem desirable to integrate nowcasting directly into each forecasting method, in practice, this is often hard to accommodate and requires considerable effort for participants in collaborative projects. We therefore split off the nowcasting from the forecasting task.

For nowcasting, we employ a chain-ladder-type method. It is based on the `simpleNowcast` method first discussed in Wolffram et al. (2023, Supplementary Section E) and has in the meantime been implemented in the R package `baselinenowcast` (Johnson et al., 2025). It combines a straightforward multiplication factor scheme with a parametric approach to estimate predictive uncertainty from past nowcast errors. Despite its simplicity, the approach showed performance comparable to more sophisticated approaches in Wolffram et al. (2023). In the present application, we need to adapt the original approach from daily to weekly data releases, which simplifies the technique because the data release and nowcast/forecast schedules now share the same frequency. The simple format of the nowcast technique allows us to handle limited or missing information on strata of the full sample that characterize our data.

3.3.1. Point nowcast

Denote by $X_{t,d}, d = 0, \dots, D$ the number of hospitalizations for week t which are added to the record with a delay of d weeks. In our applied setting, a delay $d = 0$ means that a hospitalization from the week ending on a given Sunday was already included in the data release from the following Thursday. Note that we only consider hospitalizations reported up to D weeks (in our application, $D = 4$). We now denote by

$$X_{t,\leq d} = \sum_{i=0}^{d} X_{t,i}$$

the number of hospitalizations reported for week t with a delay of at most d weeks, implying that $X_t = X_{t,\leq D}$. Conversely, for $d < D$

$$X_{t,>d} = \sum_{i=d+1}^{D} X_{t,i}$$

is the number of hospitalizations still missing after d weeks.

In the following, we write X_t etc. for a random variable, x_t for the corresponding observation, and $\hat{x}_t$ for an estimated/imputed value. The hospitalizations per week and the reporting delay, as available at a given data release time t^*, can be arranged into a *reporting triangle*, as shown in Table 1.

We consider data as available in week t^* and aim to obtain point nowcasts $\hat{x}_{t^*}, \hat{x}_{t^*-1}, \dots \hat{x}_{t^*-D+1}$, i.e., for all observations which in week t^* are still incomplete. We start by setting

$$\hat{x}_{t^*,1} = x_{t^*,0} \times \hat{\theta}_1$$

with a multiplication factor

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{N} x_{t^*-i,1}}{\sum_{i=1}^{N} x_{t^*-i,0}},$$

Table 1
Illustration of the reporting triangle for time t^* and $D = 4$. Quantities known at time t^* are shown in black and set bold for better visual distinction, yet unknown quantities are shown in gray.

week	$d = 0$	$d = 1$	$d = 2$	$d = 3$	$d = 4$	total
1	$\boldsymbol{x_{1,0}}$	$\boldsymbol{x_{1,1}}$	$\boldsymbol{x_{1,2}}$	$\boldsymbol{x_{1,3}}$	$\boldsymbol{x_{1,4}}$	$\boldsymbol{x_1}$
2	$\boldsymbol{x_{2,0}}$	$\boldsymbol{x_{2,1}}$	$\boldsymbol{x_{2,2}}$	$\boldsymbol{x_{2,3}}$	$\boldsymbol{x_{2,4}}$	$\boldsymbol{x_2}$
⋮			⋮			⋮
t^*-5	$\boldsymbol{x_{t^*-5,0}}$	$\boldsymbol{x_{t^*-5,1}}$	$\boldsymbol{x_{t^*-5,2}}$	$\boldsymbol{x_{t^*-5,3}}$	$\boldsymbol{x_{t^*-5,4}}$	$\boldsymbol{x_{t^*-5}}$
t^*-4	$\boldsymbol{x_{t^*-4,0}}$	$\boldsymbol{x_{t^*-4,1}}$	$\boldsymbol{x_{t^*-4,2}}$	$\boldsymbol{x_{t^*-4,3}}$	$\boldsymbol{x_{t^*-4,4}}$	$\boldsymbol{x_{t^*-4}}$
t^*-3	$\boldsymbol{x_{t^*-3,0}}$	$\boldsymbol{x_{t^*-3,1}}$	$\boldsymbol{x_{t^*-3,2}}$	$\boldsymbol{x_{t^*-3,3}}$	$x_{t^*-3,4}$	x_{t^*-3}
t^*-2	$\boldsymbol{x_{t^*-2,0}}$	$\boldsymbol{x_{t^*-2,1}}$	$\boldsymbol{x_{t^*-2,2}}$	$x_{t^*-2,3}$	$x_{t^*-2,4}$	x_{t^*-2}
t^*-1	$\boldsymbol{x_{t^*-1,0}}$	$\boldsymbol{x_{t^*-1,1}}$	$x_{t^*-1,2}$	$x_{t^*-1,3}$	$x_{t^*-1,4}$	x_{t^*-1}
t^*	$\boldsymbol{x_{t^*,0}}$	$x_{t^*,1}$	$x_{t^*,2}$	$x_{t^*,3}$	$x_{t^*,4}$	x_{t^*}

obtained from N preceding rows of the triangle. Here, the user chooses the estimation window size $N < t^*$ to restrict the estimation to relatively recent data. In practice, we use $N = 15$, implying that snapshots from at least the last 15 weeks are needed. Following the same principle, we compute

$$\hat{\theta}_2 = \frac{\sum_{i=2}^{N} x_{t^*-i,2}}{\sum_{i=2}^{N} x_{t^*-i,\leq 1}}$$

and use it to impute

$$\hat{x}_{t^*,2} = \hat{x}_{t^*,\leq 1} \times \hat{\theta}_2$$
$$\hat{x}_{t^*-1,2} = x_{t^*-1,\leq 1} \times \hat{\theta}_2.$$

Here, we use the $\hat{x}_{t^*,1}$ imputed in the first step to compute

$$\hat{x}_{t^*,\leq 1} = x_{t^*,0} + \hat{x}_{t^*,1}.$$

The same procedure is applied to all other missing values in the reporting triangle, which we fill from left to right and from bottom to top.

For $d = 0, \ldots, D-1$, we then sum over relevant entries of the imputed reporting triangle to obtain point nowcasts

$$\hat{x}_{t^*-d,>d} = \sum_{i=d+1}^{D} \hat{x}_{t^*-d,i}$$

for the hospitalizations from week $t^* - d$ that are still to be reported. Point nowcasts for the total numbers result as

$$\hat{x}_{t^*-d} = x_{t-d,\leq d} + \hat{x}_{t^*-d,>d}.$$

A slightly more formal explanation of how this relates to the estimation of a delay distribution from censored observations can be found in Wolffram et al. (2023). We note that this scheme would require some adaptations to handle zeros in the reporting triangle, but none occur in our setting.

3.3.2. Nowcast uncertainty

We now describe how to extend these point nowcasts to probabilistic nowcasts based on past nowcast errors. To this end, we need to slightly extend the notation and write

$$\hat{x}_{s^*-d}(s^*), \quad \hat{x}_{s^*-d,>d}(s^*), \quad \text{etc.}$$

for nowcasts referring to week $s^* - d$ and generated based on data as available in week s^*. As the uncertainty in the nowcasts stems only from hospitalizations yet to be added to the record, we focus on $\hat{x}_{s^*-d,>d}(s^*)$ in the following.

Again, consider the generation of nowcasts in week t^*. To quantify the prediction uncertainty we start by computing $\hat{x}_{s^*-d,>d}(s^*)$ for $s^* = t^* - D, \ldots, t^* - M$ and $d = 0, \ldots, D-1$. In practice, we use $M = 15$. Note that to perform these computations, data snapshots from at least $N + M$ (i.e., 30) past weeks are needed.

For each horizon $d = 0, \ldots, D-1$ we then assume that

$$X_{s^*-d,>d} \mid \hat{x}_{s^*-d,>d}(s^*) \sim$$
$$\text{NegBin}[\text{mean} = \hat{x}_{s^*-d,>d}(s^*) + 0.1, \text{disp} = \psi_d]$$

independently for each $s^* = t^* - D, \ldots, t^* - M$. An estimate $\hat{\psi}_d$ for the dispersion parameter is obtained via maximum likelihood inference. The addition of a small value of 0.1 serves to ensure well-definedness of the negative binomial distribution if $\hat{x}_{s^*-d,>d}(s^*) = 0$. In practice, we add a small tweak to also include partial observations from $s^* = t^* - 1, \ldots, t^* - D + 1$, see Wolffram et al. (2023) for details. Our nowcast distribution for $X_{t^*-d,>d}$ is then simply

$$\text{NegBin}[\text{mean} = \hat{x}_{t^*-d,>d}(t^*) + 0.1, \text{disp} = \hat{\psi}_d].$$

The corresponding distribution for the total count X_{t^*} results from shifting this distribution by the known count $x_{t^*-d,\leq d}$.

We note that if $x_{t,0} = 0$ for a given week, i.e., there are no initial releases, we remove the respective row from the reporting triangle. This helps catch weeks like Christmas, when data releases are paused.

We chose this straightforward methodology because it is straightforward to adapt to the particularity of the nowcasting task at hand. In practice, we encounter the problem that historical data snapshots are only available for the total SARI hospitalization incidence, but not the age-stratified time series (see Section 2.2). To nonetheless produce stratified nowcasts, we assume that the reporting delay distribution is identical across strata. The parameter estimates $\hat{\theta}_1, \ldots, \hat{\theta}_D$ are thus estimated from the pooled reporting triangles. The estimated overdispersion parameters $\hat{\psi}_0, \ldots, \hat{\psi}_{D-1}$ are likewise borrowed from the pooled fits.

3.3.3. Coupling of nowcasting and forecasting

For coupling the nowcast with the forecasting models, we propose the following model-agnostic approach to

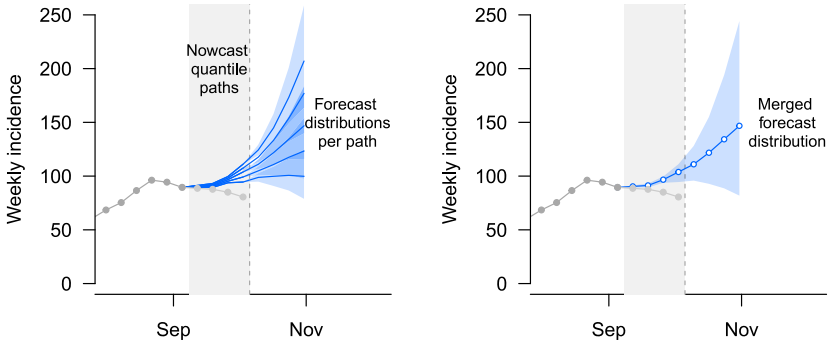

Fig. 4. Illustration of coupling between nowcasting and forecasting. A set of nowcast sample paths (blue lines in the grey shaded area) is generated. Each of these is fed into a forecasting model to obtain predictive distributions for horizons 1–4. Results are then aggregated into overall forecast distributions via a linear pool (right panel). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

propagate nowcast uncertainty into forecasts (illustrated in Fig. 4). We note that a similar scheme has been used by Brooks et al. (2018).

1. Generate nowcast distributions for horizons -3 through 0 using a separate nowcasting model. For each horizon, quantiles at $K = 39$ levels, $0.025, 0.05, \ldots, 0.95, 0.975$, are generated.
2. Translate these into 39 sample paths by assembling the predictive quantiles at identical levels for the four horizons.
3. Feed each of these paths into the employed forecasting model to generate predictive distributions for horizons 1 through 4 (depending on the method, these are samples or parametric distributions).
4. Combine these predictive distributions by aggregating samples or averaging probability mass functions with linear pooling.

Step 2 is arbitrary in a sense as the distributions our nowcasting model returns for the various horizons are purely univariate, and nothing is known about the dependence structure. However, in practice, the corrections at horizons -3 through -1 are minor and unlikely to have a significant impact on predictions. It is therefore not crucial how exactly the nowcast paths are formed. To empirically verify that this is indeed the case, we reran one of the forecasting methods discussed in the next section (`hhh4`) using randomly arranged rather than ordered paths. Similarly, we assess whether Vincentization, rather than the linear pool, yields comparable results, as the latter is also applicable to models that produce only quantile forecasts. Results on these variations are available in Section 4.2.3.

3.4. Forecasting methods

As an individual pathogen does not cause SARI, it is not straightforward to model its dynamics mechanistically using classic compartmental (SIR-type) models. However, the SARI indicator is characterized by strong autocorrelation and, at least until the COVID-19 pandemic, relatively stable seasonal patterns. It is therefore common to employ non-mechanistic statistical and machine learning models to such indicators (e.g., Albrecht et al. 2024, Mathis et al. 2024). In the following, we present a suite of such approaches. While the array of available modelling options is vast, we selected options that reflect a natural progression in complexity from parsimonious statistical modeling to "classic" machine learning (in our case, gradient boosting) and ultimately to a deep learning approach. Two of our approaches also exploit multivariate patterns across age groups (such as respiratory diseases often spreading from younger to older age groups) and information in auxiliary data streams.

3.4.1. Endemic-epidemic modeling: hhh4

The *endemic-epidemic* or `hhh4` model (after the associated function in the R package `surveillance`, Meyer et al. 2017) is a seasonal count time series model tailored for infectious disease surveillance data. It has previously been used to predict the incidence of numerous diseases, including norovirus disease (Bracher & Held, 2022), visceral leishmaniasis (Nightingale et al., 2020), and COVID-19 (Robert et al., 2024). While, in principle, it can reflect dependence structures across space or age groups, in our setting, a simple univariate formulation for each stratum proved most robust. Denoting the incidence value (as absolute count value) in week t by X_t, the model is then defined as

$$X_t \mid \text{past} \sim \text{NegBin}(\text{mean} = \lambda_t, \text{disp} = \psi) \tag{1}$$

$$\lambda_t = \nu_t + \phi_t \times \sum_{d=1}^{D} w_d X_{t-d}.$$

Here, the negative binomial distribution is parameterized by its mean λ_t and an overdispersion parameter ψ. Following (Bracher & Held, 2022), we use geometrically decaying weights w_d, while accounting for yearly seasonal variation via time-varying parameters. In the model for the pooled time series, we used the standard formulation

$$\nu_t = \alpha^{(\nu)} + \beta^{(\nu)} \times \sin(2\pi t/52.25) + \gamma^{(\nu)} \times \cos(2\pi t/52.25)$$
$$\phi_t = \alpha^{(\phi)} + \beta^{(\phi)} \times \sin(2\pi t/52.25) + \gamma^{(\phi)} \times \cos(2\pi t/52.25).$$

For the age-stratified forecasts, we further simplified this model by removing the intercept term ν_t (i.e., setting it

to zero). Even during the training period, retrospective forecasts from models including the intercept did not adapt well to changes in incidence magnitude compared to earlier seasons. Especially in the age groups 05–14 and 35–59, this led to forecasts that were poorly aligned with the preceding data points. Removing the intercept could mitigate this to a large degree.

Inference is conducted using maximum likelihood, and predictions are obtained in a simple plug-in manner. Predictive first- and second-moments can be computed analytically for all forecast horizons (Bracher & Held, 2022), and matching negative binomial distributions are used to obtain quantiles.

The model fits are updated each week using all available historical data (or, in a sensitivity analysis, excluding seasons strongly affected by the COVID-19 pandemic). Note that this also includes the corrected data points generated in the nowcasting step (see Section 3.3). Unlike the methods described in the two following subsections, no validation set is required, meaning that the distinction between the green and blue sections in Fig. 2 is not relevant here. No additional data inputs are used other than the SARI incidences.

As an additional time series benchmark that builds on related work on COVID-19 case numbers by Agosto et al. (2021), we apply a log-linear Poisson autoregressive model. In its original form, which we refer to as `Agosto1`, it is defined as

$$X_t \mid \text{past} \sim \text{Pois}(\lambda_t)$$
$$\log(\lambda_t) = \nu + \phi \times \log(X_{t-1} + 1) + \theta \times \log(\lambda_{t-1}).$$

We also propose and use an extended version of `Agosto1` where we combine a conditional negative binomial distribution as in (1), with the log-linear mean structure

$$\log(\lambda_t) = \nu + \phi \times \log(X_{t-1} + 1) + \theta \times \log(\lambda_{t-1}) + \beta \times \sin(2\pi t/52.25) + \gamma \times \cos(2\pi t/52.25),$$

thus accounting for seasonality. We refer to this model as `Agosto2` in the following. We fitted both model variants using the R package `tscount` (Liboschik et al., 2017).

3.4.2. Gradient boosting: LightGBM

`LightGBM` (Light Gradient Boosting Machine) is a gradient boosting framework designed for high-performance machine learning tasks (Ke et al., 2017). It builds decision tree ensembles sequentially, where each tree corrects the errors of the previous ones, enabling the model to capture complex patterns in the data. Its ability to efficiently handle large datasets, categorical variables, and missing values makes it versatile for a wide range of applications. In time series forecasting, `LightGBM` can effectively model relationships within multivariate data and incorporate exogenous variables. In the M5 forecasting competition (Makridakis et al., 2022), the model ranked among the top performers for predicting retail sales across multiple products and stores.

For our analysis, the model was retrained each week using available historical data and implemented in a multivariate fashion, allowing simultaneous prediction of all targets (i.e., across different age groups and the national level). Weekly ARI numbers (see Supplement B) were included as a covariate. In addition to the lagged values of these two time series (covering the previous eight weeks), the calendar week and the month of the subsequent week were incorporated as input features. The last few observations that would remain incomplete in a real-time setting were excluded from the training process. They were subsequently replaced by nowcast paths to compute the forecasts as described in Section 3.3.

Concerning hyperparameter selection, we adopted a two-stage strategy, which we implemented and recorded in the experiment tracking system *Weights and Biases* by Biewald 2020. In the first stage, we performed a random search to efficiently explore the high-dimensional parameter space and identify regions associated with good predictive performance. This approach enabled us to circumvent exhaustive evaluation of unpromising parameter combinations and to concentrate subsequent analyses on a more relevant subset. In the second stage, we conducted a systematic grid search over the refined hyperparameter ranges documented in Supplementary Table S1 to evaluate the most promising configurations thoroughly. The inclusion of the ARI covariate and the use of data from the COVID-19 period (see also Section 3.5) were part of the hyperparameter tuning, which resulted in the inclusion of both. The best-performing configurations are summarized in Supplementary Table S2. To reduce computational requirements, the model was trained once on the training dataset and evaluated across all dates in the validation period (highlighted in green and blue in Fig. 2). Due to the non-deterministic nature of the training process, we trained with ten different random seeds. We averaged the forecasts from these models (i.e., the predictive quantiles at each level) to obtain more robust results.

3.4.3. Deep learning model: TSMixer

The `TSMixer` architecture, as introduced in Chen et al. (2023), is a fully connected neural network specifically designed for time series forecasting. It utilizes a sequential mixing layer strategy that enables the model to capture both temporal dependencies and cross-feature interactions. As illustrated in Fig. 5, the mixing layers are applied sequentially: first across the time dimension to model temporal patterns and then across the feature dimension to capture relationships between different variables. This approach allows the model to learn complex, nonlinear relationships within the time series data. Compared to transformer-based models, `TSMixer` often exhibits a simpler architecture, making it more computationally efficient and easier to train. Despite its relative simplicity, `TSMixer` has demonstrated competitive performance across a wide range of time series forecasting benchmarks, suggesting that its sequential mixing-layer strategy is a practical approach for modeling temporal data. The model's ability to handle multivariate time series, as well as its potential to incorporate exogenous variables, makes it a versatile tool for a range of time series forecasting applications, including infectious disease forecasting in our setting.

The implementation, tuning, and training scheme follows that of `LightGBM` as described in the previous

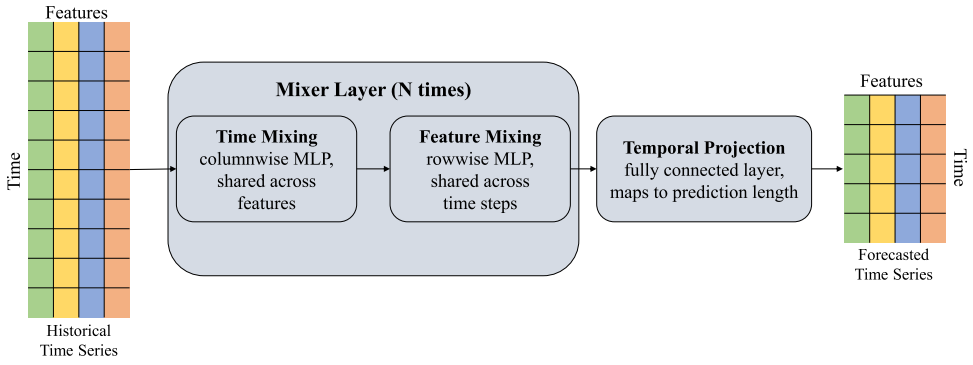

Fig. 5. Illustration of the `TSMixer` architecture, which is designed by stacking multi-layer perceptrons (MLPs). The mixing layers are applied repeatedly across time and feature dimensions to model both temporal patterns and interdependencies.

subsection. The optimized hyperparameter settings are summarized in Supplementary Table S3. As a relevant difference, we note that for `TSMixer`, hyperparameter tuning suggested removing the ARI input feature.

3.5. Variations of component models

Applying the models described in the previous sections requires many analytical choices, especially for the more complex `LightGBM` and `TSMixer` approaches. The primary settings were chosen by detailed hyperparameter tuning (Supplementary Tables S1–S3). For three aspects we consider particularly interesting, however, we report results based on alternative specifications. Firstly, we vary whether and how nowcasts are fed into the forecasting models (see Section 4.2.3 for details). Secondly, we vary how seasons affected by the acute phase of the COVID-19 pandemic are handled (see classification in Fig. 2), either including or excluding them from model fitting. Lastly, we assess the extent to which the addition of auxiliary data on acute respiratory infections (ARI) improves predictive performance. As mentioned before, while hyperparameter tuning for `LightGBM` indicated that the ARI covariate should be included, the opposite was true for `TSMixer`. The primary specifications of the two models thus differ.

3.6. The mean ensemble and reference models

For the `Ensemble`, the predictive quantiles were obtained as the arithmetic means of the individual forecasts' quantiles from the member models (`LightGBM`, `TSMixer`, and `hhh4`). This direct approach, also referred to as *Vincentization*, has been widely studied and employed in both statistics (Genest, 1992; Grushka-Cockayne et al., 2017) and machine learning (Shchur et al., 2023). We favor it over other methods, such as the linear pool, which are not applicable when only a few predictive quantiles are available. As the present analysis serves as a blueprint for a collaborative platform with quantile-based submissions (see Section 5), we work with this constraint and thus opt for the Vincentization approach. We note that a possible extension is weighted ensemble averaging, which has been explored previously for nowcasts (Amaral et al., 2025) and forecasts (e.g., Tsang et al. 2024). In particular, adaptive stacking techniques may help account for temporal variation in model performance (McAndrew & Reich, 2021).

We note that while ensemble models have often been found to outperform their individual members (e.g., Cramer et al. 2022) and are thus generally considered superior to individual models, this is not a mathematical necessity. For many combinations of ensembling procedures and scoring rules, however, results imply that the ensemble will always beat the *average* of the scores achieved by its member models; the case of the Vincentization ensemble and weighted interval score is covered by Grushka-Cockayne et al. (2017).

To put the performance of the different models into perspective, we apply two simple reference models.

- `Persistence` is an adaptation of a last-observation-carried-forward prediction to our setting with reporting delays. The predictive mean for horizons 1 through 4 is obtained as the predictive mean of the nowcast distribution at horizon 0. A predictive distribution is obtained as a negative binomial distribution with this mean value, and a dispersion parameter estimated via maximum likelihood from the 15 most recent pairs of predictive means and observations (all obtained using the respective previous data snapshots).
- `Historical` is a simplistic model that considers only past seasonal patterns. A predictive distribution for a given calendar week is obtained by collecting all available historical values for that week and the two neighboring weeks, and then fitting a negative binomial distribution.

Note that the reference models are not included in the mean ensemble.

4. Results

4.1. Visual and qualitative inspection of nowcasts and forecasts

Before turning to a formal evaluation summarizing overall performance in the next section, we provide an explorative graphical assessment of nowcasts and forecasts. Fig. 6 shows nowcasts and forecasts for the total

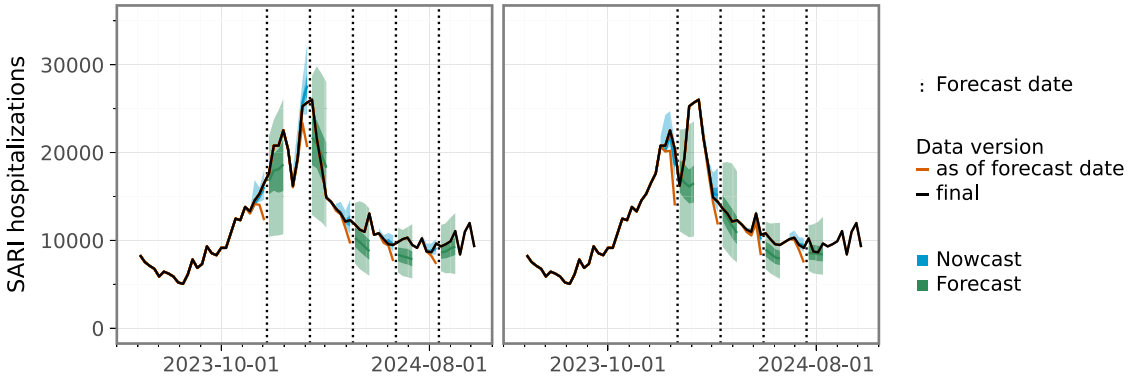

Fig. 6. Selected nowcasts and ensemble forecasts for the total SARI hospitalization incidence (pooled across age groups) at different forecast times. To avoid overplotting, we show the time series twice, overlaying it with predictions issued at different times in the two panels. Figures covering all forecast dates are available in Supplementary Figure S4.

SARI hospitalization incidence (pooled across age groups) issued by the `Ensemble` at nine different time points. To avoid overplotting, we use two separate panels and display the remaining time points in a set of Supplementary Figures (S4). A detailed illustration of the nowcasts is shown in Supplementary Figure S5. In Fig. 6, most nowcasts (blue) are closely aligned with the completed data versions (black), but in some cases discrepancies remain (e.g., for the second nowcast in the left panel). The nowcasting also successfully prevents forecasts from following spurious downward trends resulting from reporting delays. Forecasts are mostly well-aligned with the later-observed trends, except for the first weeks of 2024 (see the right panel). Here, the ensemble prediction implies that the peak has already occurred, failing to predict the second and higher peaks. Such double peaks in close succession did not happen in any of the previous years, making this aspect hard to predict solely from data. The uncertainty intervals of nowcasts and forecasts are of adequate width to nonetheless cover the observed values in most instances. Especially around the peak, however, they become very wide, making forecasts less informative in these periods.

On average over time, all models, except the `Persistence` baseline, capture the qualitative seasonal patterns well. This can be seen from the respective values of the rank-based RGA measure in Table 2. All models except `Persistence` achieve RGA values close to the optimal value of one, with only minor differences between models. Our interpretation of these results is that the seasonal structure of the SARI time series is sufficiently stable to make it reasonably easy to get the rank structure across weeks right. The more challenging task lies in predicting the magnitude of the SARI curve, which can vary from year to year.

Selected predictions from individual models across age groups are displayed in Fig. 7. As discussed in Section 2.1, age group 15–34 displayed unusual patterns in the 2023/24 season. Unlike in previous years, incidence remained relatively high throughout the spring and summer. The `LightGBM` model struggles to adapt to this difference and continues to predict a decline towards the usual levels (a similar pattern occurs with `TSMixer`). The

Table 2
RGA values for the total SARI hospitalization incidence (pooled across age groups, separately per prediction horizon). This corresponds to sets of $n = 48$ predictions and observations. As the point predictions, we used the predictive medians from our different models.

Model	Horizon 1	Horizon 2	Horizon 3	Horizon 4
`Ensemble`	0.970	0.961	0.951	0.951
`LightGBM`	0.967	0.952	0.935	0.926
`TSMixer`	0.959	0.939	0.923	0.920
`hhh4`	0.970	0.957	0.948	0.940
`Persistence`	0.964	0.940	0.912	0.883
`Historical`	0.956	0.960	0.962	0.966

`hhh4` model, with its simple autoregressive structure, is better able to handle this shift in magnitude. The difficulties of `LightGBM` and `TSMixer` are also inherited by the `Ensemble`. Similar patterns are also found for age group 05–14, and to a lesser degree for ages 35–59, while the remaining age groups have more typical seasonal courses. However, Fig. 7 also illustrates some strengths of `LightGBM` and `TSMixer`, particularly at the national level (00+) and for older age groups (e.g., 80+). These models accurately capture the sharp decline following the second peak, whereas `hhh4` tends to produce more pessimistic forecasts.

4.2. Formal forecast evaluation

4.2.1. Aggregate-level nowcasts and forecasts

We complement the visual assessment with a more formal evaluation of forecast calibration and score-based performance. Fig. 8 summarizes the performance for the total hospitalization incidence (pooled across age groups). Average WIS (across forecast dates) and the coverage fractions for the 50% and 95% prediction intervals are displayed stratified by nowcast/forecast horizon. Surprisingly, average scores increase with the horizon (i.e., performance decreases). For horizons 1 through 4, all models outperform the `Persistence` and `Historical` baseline models (except for TSMixer at horizon 1). The `Ensemble` outperforms all individual models at all horizons, but the margin over `LightGBM` and `hhh4` is slim at short horizons (and indeed, most score differences are not statistically

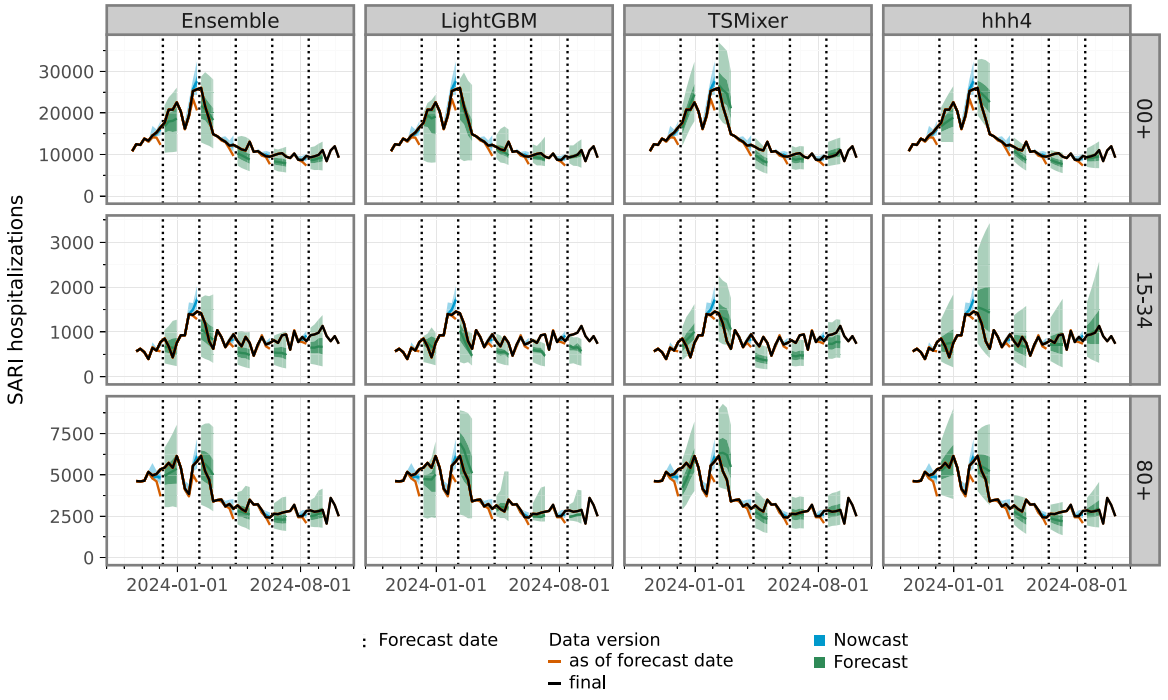

Fig. 7. Selected nowcasts and forecasts for the aggregate level 00+ and age groups 15–34 and 80+. To avoid overplotting, predictions for only five forecast times are shown. Figures covering all forecast dates are available for the `Ensemble` in Supplementary Figure S4.

significant; see below). Interestingly, for horizon 4, this flips, and the `TSMixer` model achieves performance close to the ensemble. The decomposition of the WIS indicates that `LightGBM` and `TSMixer` tend to underpredict, and that the ensemble inherits this tendency (this seems to be driven by the fact that the second peak was not anticipated, as well as the unusually high incidences of some age groups late in the season; see previous subsection). The `hhh4` and nowcasting models have more balanced components.

A summary plot aggregating results across horizons is available in Supplementary Figure S6 (left panel). While the `Ensemble` again has a little edge, the three-member models `LightGBM`, `TSMixer`, and `hhh4` are roughly on par. Concerning the interval coverage rates (bottom panel in Fig. 8), all models apart from the `Historical` baseline achieve close-to-nominal coverage. Displays of calibration and average scores stratified by quantile level are available in Supplementary Figures S10 and S11. The relative performance of models is consistent across quantile levels, except for `LightGBM`, which shows a drop in performance at higher quantiles.

As detailed in Supplementary Figure S12, Diebold–Mariano tests (cited in Diebold and Mariano 2002, implemented in Leeuwenburg et al. 2024) indicate that the observed score differences are mostly not significant. While the `Ensemble` has significantly better performance than the `Historical` baseline at all horizons, differences to the `Persistence` baseline are only significant for certain combinations of competing model and horizon (e.g., `hhh4` and `LightGBM` at horizon 2). This may reflect that the DM test often has low power in small to moderate sample sizes, especially when tests are performed against naïve baseline models (Coroneo & Iacone, 2025).

The performance of the variations `Agosto1` and `Agosto2` of the `hhh4` model is displayed in Supplementary Figure S9. The modest results for the simplistic `Agosto1` indicate that it is important to account for seasonality and overdispersion. Specifically, the Poisson assumption of `Agosto1` yields overly narrow prediction intervals, leading to its performance even falling behind both baseline methods. When augmenting the model with a negative binomial distribution and sine/cosine terms for seasonality (`Agosto2`), performance is practically equivalent to that of `hhh4`.

4.2.2. Age-stratified nowcasts and forecasts

Fig. 9 summarizes average results for age-stratified nowcasts and forecasts. The results for average WIS are broadly consistent with those discussed in the previous section, with the ensemble again performing best across horizons and the individual models outperforming the baseline models in almost all cases. The `LightGBM` and `TSMixer` models again tend to underpredict, while the `hhh4` model features the most dispersed predictions.

The WIS stratified by age group (and aggregated by horizon), depicted in Fig. 10, reveals that the aforementioned downward bias in `LightGBM` and `TSMixer` primarily originates from the age groups 05–14, 15–34, and 35–59. This can be attributed to the unusually high SARI incidence during the evaluation period (Fig. 2), which did not follow the typical seasonal decline, as discussed previously. The score-based evaluation also confirms that the `hhh4` model performs particularly well in these age

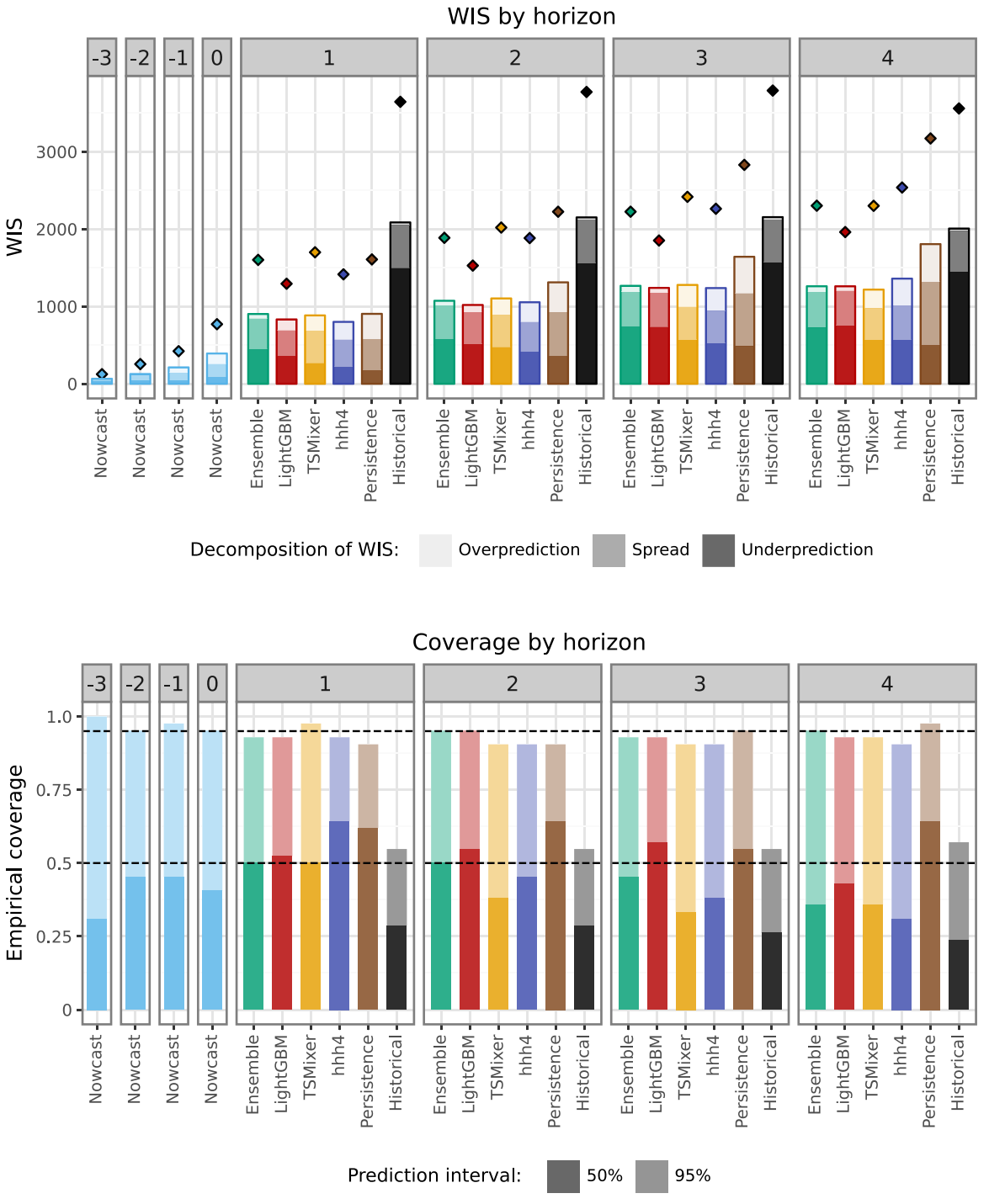

Fig. 8. Top: Average WIS values (bars) and absolute errors (diamonds) achieved by different models for the total SARI hospitalization incidence. The average WIS scores are decomposed into components for overprediction, underprediction, and forecast spread. Bottom: Empirical coverage rates of 50% and 95% prediction intervals.

groups (in the 15–34 group, even slightly outperforming the `Ensemble`). By contrast, the machine learning approaches had an edge in forecasting older age groups, potentially because they could leverage trends in younger age groups as leading indicators for older ones.

In terms of interval coverage (bottom panel of Fig. 9), we observe that the nowcasts for horizons −1 and 0 are considerably overconfident. This is likely a consequence of the fact that only a few historical snapshots of age-stratified data were available, meaning that stratified

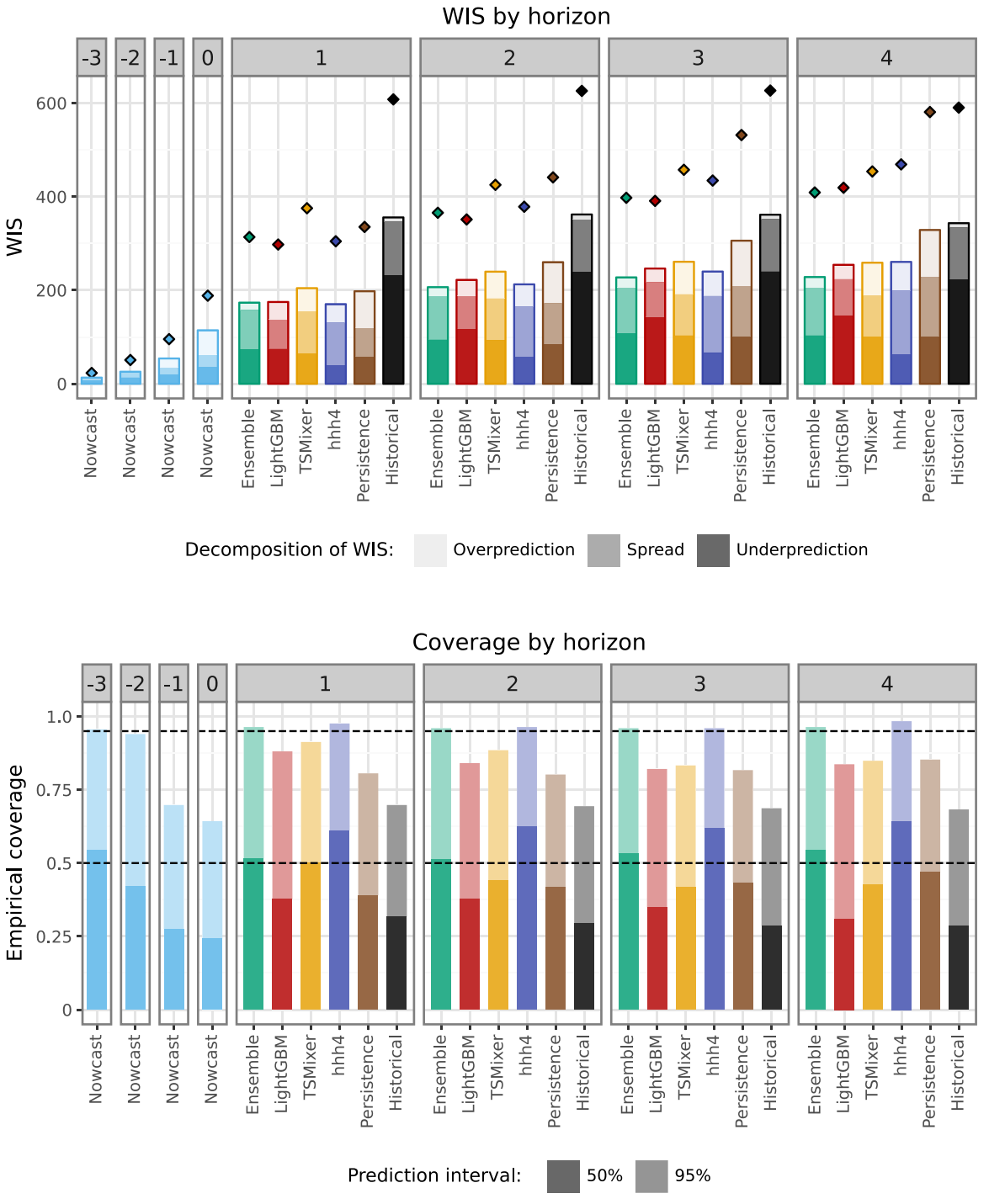

Fig. 9. Top: Average WIS values (bars) and absolute errors (diamonds) achieved by different models for the age-stratified SARI hospitalization incidence. The average WIS scores are decomposed into components for overprediction, underprediction, and forecast spread. Bottom: Empirical coverage rates of 50% and 95% prediction intervals.

nowcasts had to be based on aggregate-level snapshots (see Section 3.3). The forecasts from the `LightGBM` and to a lesser degree `TSMixer` models are somewhat overconfident, too. This is not surprising given that the forecasting models take the nowcast as an input. Remarkably, the `Ensemble` forecast is well-calibrated across horizons and interval levels. This can be explained by the fact that, when using Vincentization, the ensemble prediction intervals have an average length equal to the member intervals. If the ensemble intervals are centered around

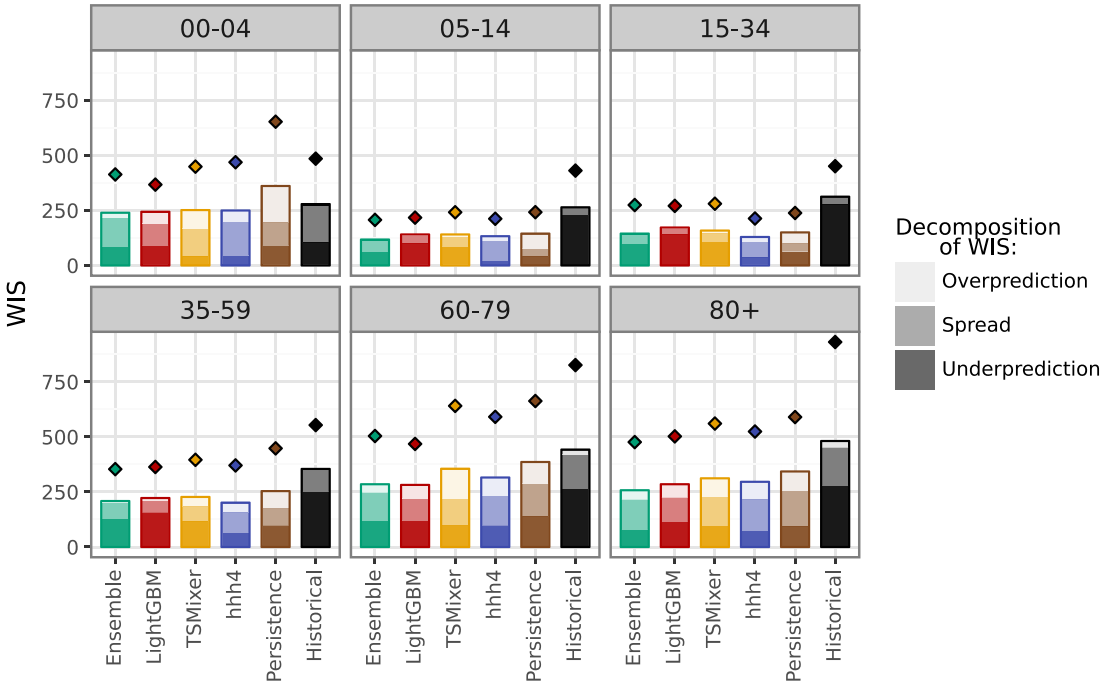

Fig. 10. Average WIS (bars) and absolute errors (diamonds) by age group, aggregated over forecast dates and horizons. Average scores are decomposed into components for overprediction, underprediction, and forecast spread.

a more accurate central tendency (as is often the case), interval coverage rates will tend to increase.

4.2.3. Integration of nowcasts and forecasts

For each forecasting method, we investigate the impact of integrating nowcasts into forecasts and assess the performance of the chosen implementation approach. Thus, instead of including nowcast distributions as described in Section 3.3, we apply three alternative strategies.

(i) Firstly, we simply ignore the delay problem and use uncorrected incomplete data points to initialize our forecasting models ("Naive").
(ii) Secondly, we discard the last available observation and use only largely stable observations, as is common in the literature (Paireau et al., 2022). We still apply the nowcasting procedure to the previous weeks, but this makes little practical difference ("Discard").
(iii) Lastly, we base forecasts on the final versions of the latest data points, i.e., assess how much forecasts would improve if the reporting system were free of delays. This is a hypothetical setting and not an approach that could be applied in real time ("Oracle").

Fig. 11 summarizes the performance for the total SARI hospitalization incidence when using the four considered ways of handling recent data points. Our proposed method of including the latest data point with a nowcast correction ("Coupling") yields improvements over using uncorrected data ("Naive") or discarding this data point ("Discard"). This holds especially for short horizons, where forecast initialization is most relevant. In fact, for the `hhh4` model, the "Discard" version even works slightly better for horizons 3 and 4. Somewhat surprisingly, when providing forecast models with the final values of recent data points ("Oracle") rather than nowcasts, performance does not always improve. While `hhh4` does, the other models show minor performance deterioration at some horizons. A possible explanation is that initializing the models `LightGBM` and `TSMixer` with a nowcast distribution rather than the correct value increases forecast dispersion, thereby improving calibration. Corresponding results for age-stratified predictions are shown in Supplementary Figure S8 and are in good agreement with the aggregate-level results.

To assess whether our choice to obtain different sample paths by ordering the quantiles per week affects the results, we reran the "coupling" approach for `hhh4` using randomly arranged paths. As shown in Supplementary Figure S9, the resulting scores remain virtually unchanged (model denoted `hhh4-Shuffle`). Moreover, we assessed performance when using quantile averaging rather than the linear pool to combine forecasts from different nowcast paths. This is relevant for models that only issue predictive quantiles rather than full predictive distributions (the linear pool is obviously not available in this case), as shown in Supplementary Figure S9 (model denoted `hhh4-Vincentization`), this, too, makes little difference in practice.

4.2.4. Handling of the acute phase of the COVID-19 pandemic and inclusion of ARI data

Finally, we study how the handling of the COVID-19 period and auxiliary data on ARI consultations affect

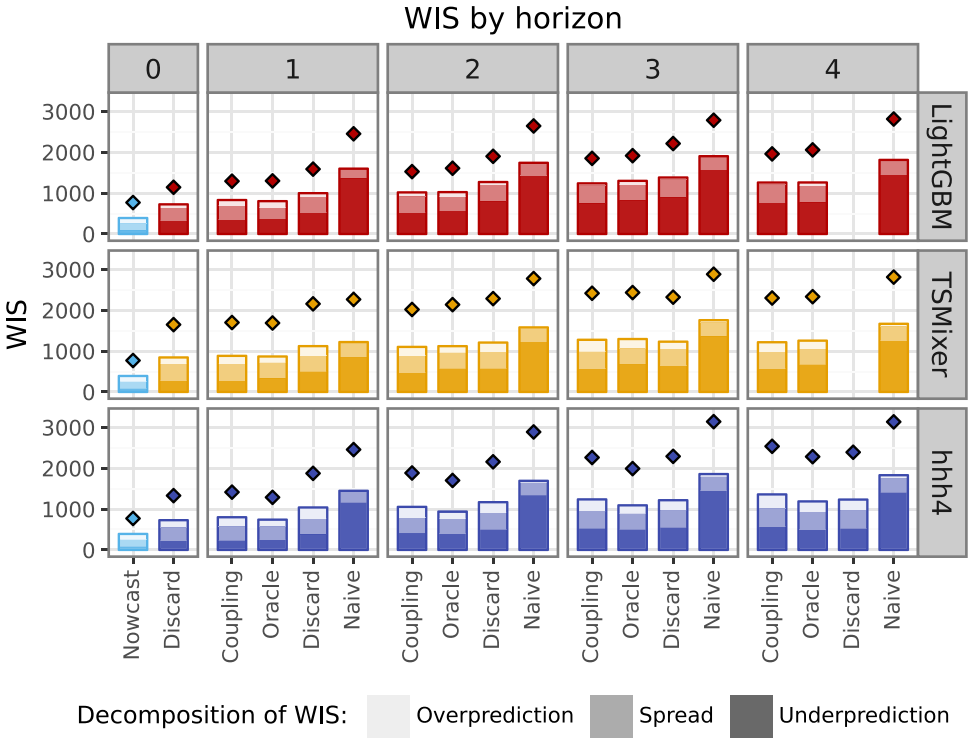

Fig. 11. Comparison of forecast performance on the aggregate level resulting from different strategies to handle incomplete recent data. "Coupling" is our main approach described in Section 3.3, i.e., feeding the full nowcast into forecasting models. "Discard" corresponds to discarding the most recent (i.e., most incomplete) data point and treating it like an additional value to be predicted. "Naive" uses the time series as is (with yet incomplete values). "Oracle" is a hypothetical setting where the final versions of the most recent data points are used. It thus enables us to assess the impact of reporting delays on forecast quality. As in previous figures, the bars show WIS, and the diamonds show absolute errors.

the predictive performance of each forecasting method (see Section 3.5). In each setting, the ML models were retrained with hyperparameters optimized explicitly for that setting. As before, due to the non-deterministic nature of the training process, we trained with 10 different random seeds and averaged the forecasts from these models (i.e., the predictive quantiles at each level) to obtain more robust results.

Supplementary Figure S9 summarizes the effect of excluding data from the COVID-19 period in the training set as well as from using ARI incidences as an auxiliary data stream for `LightGBM` and `TSMixer`. Discarding the data from the COVID-19 period led to a slight deterioration in performance for `hhh4` and `LightGBM`. The `TSMixer` model was very poorly behaved when applied to a reduced data set excluding the COVID-19 period, indicating that our full time series may already be near the lower end of this method's data requirements.

The inclusion of the auxiliary time series on outpatient consultations for ARI was not beneficial; see Supplementary Figure S9. For `LightGBM`, where hyperparameter tuning suggested including the ARI data, excluding it yielded essentially identical performance on the test set. In the case of `TSMixer`, where the covariate was excluded based on the hyperparameter tuning, the performance drop when including it was rather substantial in the test set.

5. Discussion and conclusions

We presented and evaluated a multi-model system for nowcasting and short-term forecasting of hospitalizations from severe acute respiratory infections (SARI) in Germany. We addressed this in a modular fashion, generating nowcasts in a separate step and subsequently feeding them into the forecasting models. For short forecast horizons, this led to improvements over a more straightforward approach that used the most recent data points uncorrected or simply discarded them. Compared with simple baseline forecasting methods, the three models considered showed improved performance across lead times, though the observed differences were not consistently statistically significant. Similar to previous efforts, we found that a combined ensemble prediction improved upon individual models. Forecasts were generally well-calibrated in terms of interval coverage fractions, but in some models, as well as the ensemble, we observed noteworthy biases in some age groups. In these instances, the machine learning models `LightGBM` and `TSMixer` seemed to overfit historical patterns, while the more straightforward statistical approach `hhh4` performed better. In age groups where the seasonal course was closer to historical patterns, however, this model had weaker relative performance.

The good probabilistic calibration of almost all considered models represents a marked difference from results achieved in recent years for COVID-19 cases or deaths (see e.g., Bracher et al. 2021, Cramer et al. 2022). This is undoubtedly not due to a sudden improvement in forecasting capacities, but due to the higher predictability of seasonal disease dynamics. Unlike in COVID-19 forecasting, social dynamics and intervention measures were unlikely to be major drivers during the test period. Also, reporting practices were considerably more stable than for most COVID-19 indicators.

Our analyses of forecast performance across horizons and age groups indicate that our three stand-alone models have differing strengths and weaknesses. This *ensemble diversity* is often considered a key feature of good ensemble performance (DelSole et al., 2014). Especially during the COVID-19 pandemic, collaborative forecasting projects featured considerably more models (the largest effort likely being Cramer et al. 2022 with more than 100 models). This level of effort is unrealistic and undesirable outside of times of major crisis. How many models need to be run to achieve robust ensemble performance is currently under research. Fox et al. (2024) recommend using four to seven models and find that the gain from additional models diminishes quickly. In future operational use of our system (see below), two more independently run models will be included for SARI hospitalizations. We hope this will further enhance the ensemble's robustness, while keeping the required effort at a sustainable level.

In the present work, we consider only SARI hospitalizations at the national level, with age stratification. Unfortunately, it is currently not feasible to analyze these data at a regional level. Coverage by sentinel hospitals varies considerably across the 16 German states, with several of them not covered at all (Buda et al., 2017). State-level estimates are thus not released by RKI. Also, it was not feasible to run a validation on a longer test period, which would have strengthened the generalizability of our results. The reason is that vintage data snapshots were unavailable for earlier time periods, preventing us from studying the integration of nowcasting and forecasting during those periods.

The presented work serves as a blueprint for the *RESPINOW Hub* (http://respinowhub.de/), an operational disease nowcasting and forecasting system. Launched in Fall 2024, the Hub covers multiple prediction targets (including the outpatient ARI consultation incidence discussed in Section 2.3 and mandatory case reporting of several respiratory diseases). A follow-up study on the real-time performance of different models across indicators has been preregistered (Bracher & Wolffram, 2024).

This follow-up study will enable us to address one of the significant weaknesses of the present project: the risk of hindsight bias. While we made considerable efforts to manage historical data versions correctly and avoid using data that would not have been available in real time, the development and evaluation of prediction models are iterative processes. Implicitly, some knowledge of the test set's characteristics may thus have diffused into our forecasting approaches.

Another limitation of our approach is that we consider only one aggregate syndromic indicator, thereby limiting the applicability of mechanistic (SIR-type) models. These may be more proficient at predicting tipping points due to the depletion of susceptibles, even when seasonal patterns differ from previous years. A promising recent development is that the Robert Koch Institute has started releasing stratified data on SARI hospitalizations caused by COVID-19, seasonal influenza, and RSV. In future years, pathogen-specific mechanistic models can thus be applied. Such a stratified approach may ultimately also lead to improved forecasts of the total SARI hospitalization incidence.

Data and code availability

All results in this paper can be reproduced using the publicly available replication package at https://github.com/dwolffram/replication-sari-forecasting that contains all data and code.

CRediT authorship contribution statement

Daniel Wolffram: Writing – original draft, Visualization, Validation, Software, Formal analysis, Data curation. **Johannes Bracher:** Writing – original draft, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Melanie Schienle:** Writing – original draft, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Acknowledgments

This research was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) via the project RESPINOW (grant number: HZI MV2021-012, NMI FKZ031L0298B). Melanie Schienle also acknowledges support by the Klaus Tschira Foundation. Johannes Bracher was moreover supported by the German Research Foundation (DFG), project 512483310.

We would like to thank Sam Abbott, Davide Hailer, and Jan van de Kassteele for helpful discussions.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Melanie Schienle, Johannes Bracher reports financial support was provided by Federal Ministry of Education and Research Bonn Office. Melanie Schienle reports a relationship with Heidelberg Institute for Theoretical Studies that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2026.01.001.

References

Agosto, Arianna, Campmas, Alexandra, Giudici, Paolo, & Renda, Andrea (2021). Monitoring COVID-19 contagion growth. *Statistics in Medicine*, *40*(18), 4150–4160.

Albrecht, S, Broderick, D, Dost, K, Cheung, I, Nghiem, N, Wu, M, Zhu, J, Poonawala-Lohani, N, Jamison, S, Rasanathan, D, Huang, S, Trenholme, A, Stanley, A, Lawrence, S, Marsh, S, Castelino, L, Paynter, J, Turner, N, McIntyre, P, ... Wicker, JS (2024). Forecasting severe respiratory disease hospitalizations using machine learning algorithms. *BMC Medical Informatics and Decision Making*, *24*(293).

Amaral, André Victor Ribeiro, Wolffram, Daniel, Moraga, Paula, & Bracher, Johannes (2025). Post-processing and weighted combination of infectious disease nowcasts. *PLoS Computational Biology*, *21*(3), 1–24.

Beesley, Lauren J., Osthus, Dave, & Del Valle, Sara Y. (2022). Addressing delayed case reporting in infectious disease forecast modeling. *PLoS Computational Biology*, *18*(6), 1–26. http://dx.doi.org/10.1371/journal.pcbi.1010115.

Biewald, Lukas (2020). Experiment tracking with weights and biases. URL: https://www.wandb.com/. Software available from wandb.com.

Bleichrodt, Amanda, Luo, Ruiyan, Kirpich, Alexander, & Chowell, Gerardo (2024). Evaluating the forecasting performance of ensemble sub-epidemic frameworks and other time series models for the 2022–2023 mpox epidemic. *Royal Society Open Science*, *11*(7), Article 240248. http://dx.doi.org/10.1098/rsos.240248.

Bracher, Johannes, & Held, Leonhard (2022). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *International Journal of Forecasting*, *38*(3), 1221–1233. http://dx.doi.org/10.1016/j.ijforecast.2020.07.002.

Bracher, Johannes, & Wolffram, Daniel (2024). Preregistration: Nowcasting and short-term forecasting of respiratory infections in Germany, 2024/25. https://osf.io/tgsem/.

Bracher, Johannes, Wolffram, Daniel, Deuschel, Jannik, Görgen, Konstantin, Ketterer, Jakob L, Ullrich, Alexander, Abbott, Sam, Barbarossa, Maria Vittoria, Bertsimas, Dimitris, Bhatia, Sangeeta, et al. (2021). A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nature Communications*, *12*(1), 5173.

Brooks, Logan C., Farrow, David C., Hyun, Sangwon, Tibshirani, Ryan J., & Rosenfeld, Roni (2018). Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS Computational Biology*, *14*(6), 1–29.

Buchholz, Udo, Lehfeld, Ann-Sophie, Tolksdorf, Kristin, Cai, Wei, Reiche, Janine, Biere, Barbara, Dürrwald, Ralf, & Buda, Silke (2023). Respiratory infections in children and adolescents in Germany during the COVID-19 pandemic. *Journal of Health Monitoring*, *8*(2), 20.

Buda, Silke, Tolksdorf, Kristin, Schuler, Ekkehard, Kuhlen, Ralf, & Haas, Walter (2017). Establishing an ICD-10 code based SARI-surveillance in Germany–description of the system and first results from five recent influenza seasons. *BMC Public Health*, *17*, 1–13.

Charniga, Kelly, Madewell, Zachary J., Masters, Nina B., Asher, Jason, Nakazawa, Yoshinori, & Spicknall, Ian H. (2024). Nowcasting and forecasting the 2022 U.S. mpox outbreak: Support for public health decision making and lessons learned. *Epidemics*, *47*, Article 100755.

Chen, Si-An, Li, Chun-Liang, Yoder, Nate, Arik, Sercan O, & Pfister, Tomas (2023). TSMixer: An all-MLP architecture for time series forecasting. arXiv preprint arXiv:2303.06053.

Coroneo, Laura, & Iacone, Fabrizio (2025). Testing for equal predictive accuracy with strong dependence. *International Journal of Forecasting*, *41*(3), 1073–1092.

Cramer, Estee Y, Ray, Evan L, Lopez, Velma K, Bracher, Johannes, Brennen, Andrea, Castro Rivadeneira, Alvaro J, Gerding, Aaron, Gneiting, Tilmann, House, Katie H, Huang, Yuxin, et al. (2022). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, *119*(15), Article e2113561119.

De Nicola, G., M., Schneble, G., Kauermann, & Berger, U. (2022). Regional now- and forecasting for data reported with delay: toward surveillance of COVID-19 infections. *AStA. Advances in Statistical Analysis*, *106*, 407–426.

DelSole, Timothy, Nattala, Jyothi, & Tippett, Michael K. (2014). Skill improvement from increased ensemble size and model diversity. *Geophysical Research Letters*, *41*(20), 7331–7342.

Diebold, Francis X., & Mariano, Robert S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Fiandrino, Stefania, Bizzotto, Andrea, Guzzetta, Giorgio, Merler, Stefano, Baldo, Federico, Valdano, Eugenio, Urdiales, Alberto Mateo, Bella, Antonino, Celino, Francesco, Zino, Lorenzo, Rizzo, Alessandro, Li, Yuhan, Perra, Nicola, Gioannini, Corrado, Milano, Paolo, Paolotti, Daniela, Quaggiotto, Marco, Rossi, Luca, Vismara, Ivan, ... Gozzi, Nicolò (2025). Collaborative forecasting of influenza-like illness in Italy: The Influcast experience. *Epidemics*, *50*, Article 100819.

Fox, Spencer J, Kim, Minsu, Meyers, Lauren Ancel, Reich, Nicholas G, & Ray, Evan L (2024). Optimizing disease outbreak forecast ensembles. *Emerging Infectious Diseases*, *30*(9), 1967.

Funk, S, Abbott, S, Atkins, BD, Baguelin, M, Baillie, JK, Birrell, P, Blake, J, Bosse, NI, Burton, J, Carruthers, J, Davies, NG, De Angelis, D, Dyson, L, Edmunds, WJ, Eggo, RM, Ferguson, NM, Gaythorpe, K, Gorsich, E, Guyver-Fletcher, G, ... Investigators, ISARIC4C (2021). Short-term forecasts to inform the response to the COVID-19 epidemic in the UK. MedRxiv. URL: https://doi.org/10.1101/2020.11.11.20220962.

Genest, Christian (1992). Vincentization revisited. *The Annals of Statistics*, 1137–1142.

Giudici, P., & Raffinetti, E. (2025). RGA: a unified measure of predictive accuracy. *Advances in Data Analysis and Classification*, *19*, 67–93.

Gneiting, Tilmann, Raftery, Adrian E, Westveld, Anton H, & Goldman, Tom (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118.

Goerlitz, Luise, Tolksdorf, Kristin, Buchholz, Udo, Prahm, Kerstin, Preuß, Ute, an der Heiden, Matthias, Wolff, Thorsten, Dürrwald, Ralf, Nitsche, Andreas, Michel, Janine, Haas, Walter, & Buda, Silke (2021). Überwachung von COVID-19 durch Erweiterung der etablierten Surveillance für Atemwegsinfektionen. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *64*(4), 1437–1588. http://dx.doi.org/10.1007/s00103-021-03303-2.

Grushka-Cockayne, Yael, Lichtendahl, Kenneth C., Jose, Victor Richmond R., & Winkler, Robert L. (2017). Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Operations Research*, *65*(3), 712–728.

Johnson, Kaitlyn E., Tang, Maria L., Tyszka, Emily, Jones, Laura, Nemcova, Barbora, Wolffram, Daniel, Ergas, Rosa, Reich, Nicholas G., Funk, Sebastian, Mellor, Jonathon, Bracher, Johannes, & Abbott, Sam (2025). Baseline nowcasting methods for handling delays in epidemiological data. MedRxiv. preprint at https://www.medrxiv.org/content/early/2025/08/15/2025.08.14.25333653.

Ke, Guolin, Meng, Qi, Finley, Thomas, Wang, Taifeng, Chen, Wei, Ma, Weidong, Ye, Qiwei, & Liu, Tie-Yan (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.

Leeuwenburg, Tennessee, Loveday, Nicholas, Ebert, Elizabeth E., Cook, Harrison, Khanarmuei, Mohammadreza, Taggart, Robert J., Ramanathan, Nikeeth, Carroll, Maree, Chong, Stephanie, Griffiths, Aidan, & Sharples, John (2024). Scores: A python package for verifying and evaluating models and predictions with xarray. *Journal of Open Source Software*, *9*(99), 6889. http://dx.doi.org/10.21105/joss.06889.

Liboschik, Tobias, Fokianos, Konstantinos, & Fried, Roland (2017). Tscount: An r package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, *82*(5), 1–51.

Ma, Long, Qiu, Zhihao, Van Mieghem, Piet, & Kitsak, Maksim (2024). Reporting delays: A widely neglected impact factor in COVID-19 forecasts. *PNAS Nexus*, *3*(6), pgae204.

Makridakis, Spyros, Spiliotis, Evangelos, & Assimakopoulos, Vassilios (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364.

Mathis, S, Webber, AE, León, TM, Murray, EL, Sun, M, White, LA, Brooks, LC, Green, A, Hu, AJ, Rosenfeld, R, Shemetov, D, Tibshirani, RJ, McDonald, DJ, Kandula, S, Pei, S, Yaari, R, Yamana, TK, Shaman, J, Agarwal, P, ... Borchering, RK (2024). Evaluation of FluSight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations. *Nature Communications*, *15*(6289).

McAndrew, Thomas, & Reich, Nicholas G. (2021). Adaptively stacking ensembles for influenza forecasting. *Statistics in Medicine*, *40*(30), 6931–6952.

Meyer, Sebastian, Held, Leonhard, & Höhle, Michael (2017). Spatio-temporal analysis of epidemic phenomena using the r package surveillance. *Journal of Statistical Software*, *77*(11), 1–55. http://dx.doi.org/10.18637/jss.v077.i11.

Nightingale, Emily S., Chapman, Lloyd A. C., Srikantiah, Sridhar, Subramanian, Swaminathan, Jambulingam, Purushothaman, Bracher, Johannes, Cameron, Mary M., & Medley, Graham F. (2020). A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India. *PLoS Neglected Tropical Diseases*, *14*(7), 1–21.

Osthus, Dave, Daughton, Ashlynn R., & Priedhorsky, Reid (2019). Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLoS Computational Biology*, *15*(2), 1–19.

Paireau, Juliette, Andronico, Alessio, Hozé, Nathanaël, Layan, Maylis, Crépey, Pascal, Roumagnac, Alix, Lavielle, Marc, Boëlle, Pierre-Yves, & Cauchemez, Simon (2022). An ensemble model based on early predictors to forecast COVID-19 health care demand in France. *Proceedings of the National Academy of Sciences*, *119*(18), Article e2103302119. http://dx.doi.org/10.1073/pnas.2103302119.

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

Reich, Nicholas G, Lessler, Justin, Funk, Sebastian, Viboud, Cecile, Vespignani, Alessandro, Tibshirani, Ryan J, Shea, Katriona, Schienle, Melanie, Runge, Michael C, Rosenfeld, Roni, et al. (2022). Collaborative hubs: making the most of predictive epidemic modeling. *American Journal of Public Health*, *112*(6), 839–842.

Robert, Alexis, Chapman, Lloyd A. C., Grah, Rok, Niehus, Rene, Sandmann, Frank, Prasse, Bastian, Funk, Sebastian, & Kucharski, Adam J. (2024). Predicting subnational incidence of COVID-19 cases and deaths in EU countries. *BMC Infectious Diseases*, *24*(article number 204).

Shchur, Oleksandr, Turkmen, Ali Caner, Erickson, Nick, Shen, Huibin, Shirkov, Alexander, Hu, Tony, & Wang, Bernie (2023). AutoGluon–TimeSeries: Automl for probabilistic time series forecasting. In Aleksandra Faust, Roman Garnett, Colin White, Frank Hutter, & Jacob R. Gardner (Eds.), *Proceedings of machine learning research*: *vol. 224*, *Proceedings of the second international conference on automated machine learning* (pp. 9/1–21). PMLR.

Sherratt, Katharine, Gruson, Hugo, Johnson, Helen, Niehus, Rene, Prasse, Bastian, Sandmann, Frank, Deuschel, Jannik, Wolffram, Daniel, Abbott, Sam, Ullrich, Alexander, et al. (2023). Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *ELife*, *12*, Article e81916.

Tolksdorf, Kristin, Haas, Walter, Schuler, Ekkehard, Wieler, Lothar H., Schilling, Julia, Hamouda, Osamah, Diercke, Michaela, & Buda, Silke (2022). ICD-10 based syndromic surveillance enables robust estimation of burden of severe COVID-19 requiring hospitalization and intensive care treatment. http://dx.doi.org/10.1101/2022.02.11.22269594.

Tsang, T. K., Du, Q., Cowling, B. J., & Viboud, C. (2024). An adaptive weight ensemble approach to forecast influenza activity in an irregular seasonality context. *Nature Communications*, *15*(862).

Wolffram, Daniel, Abbott, Sam, An der Heiden, Matthias, Funk, Sebastian, Günther, Felix, Hailer, Davide, Heyder, Stefan, Hotz, Thomas, van de Kassteele, Jan, Küchenhoff, Helmut, et al. (2023). Collaborative nowcasting of COVID-19 hospitalization incidences in Germany. *PLoS Computational Biology*, *19*(8), Article e1011394.