Post-processing and weighted combination of infectious disease nowcasts

André Victor Ribeiro Amaral[1,2], Daniel Wolffram[3,4], Paula Moraga[1], and Johannes Bracher[3,4,*]

[1] CEMSE Division, King Abdullah University of Science and Technology. Thuwal, Saudi Arabia.
[2] Current address: Department of Mathematics. Imperial College London, London, United Kingdom.
[3] Institute of Statistics, Karlsruhe Institute of Technology. Karlsruhe, Germany.
[4] Computational Statistics Group, Heidelberg Institute for Theoretical Studies. Heidelberg, Germany.
[*] Corresponding author. E-mail: johannes.bracher@kit.edu

Abstract

In infectious diseases surveillance, incidence data are frequently subject to reporting delays and retrospective corrections, making it hard to assess current trends in real time. A variety of probabilistic nowcasting methods have been suggested to correct for the resulting biases. Building upon a recent comparison of eight of these methods in an application to COVID-19 hospitalization data from Germany, the objective of this paper is twofold. Firstly, we investigate how nowcasts from different models can be improved using statistical post-processing methods as employed, e.g., in weather forecasting. Secondly, we assess the potential of weighted ensemble nowcasts, i.e., weighted combinations of different probabilistic nowcasts. These are a natural extension of unweighted nowcast ensembles, which have previously been found to outperform most individual models. Both in post-processing and ensemble building, specific challenges arise from the fact that data are constantly revised, hindering the use of standard approaches. We find that post-processing can improve the individual performance of almost all considered models both in terms of evaluation scores and forecast interval coverage. Improving upon the performance of unweighted ensemble nowcasts via weighting schemes, on the other hand, poses a substantial challenge. Across an array of approaches, we find modest improvement in scores for some and decreased performance for most, with overall more favorable results for simple methods. In terms of forecast interval coverage, however, our methods lead to rather consistent improvements over the unweighted ensembles.

Keywords: Ensemble; Forecasting; Infectious diseases; Nowcasting; Post-processing.

1 Introduction

Real-time surveillance plays a critical role in monitoring and analyzing the spread of infectious diseases, but the availability of timely and accurate data remains a challenge. The nature of data collection and reporting introduces delays, which cause recent data points to be incomplete and trends difficult to assess. Statistical nowcasting methods can be employed to predict by how much recent values will be corrected upwards.

Such methods have been extensively employed in various infectious disease settings, including dengue (Codeco et al., 2018; Bastos et al., 2019; Beesley et al., 2022), HIV (Cox and Medley, 1989) and outbreaks of gastrointestinal diseases (Höhle and an der Heiden, 2014). During the COVID-19 pandemic, the topic received increased attention (Greene et al., 2021; Günther et al., 2021; Seaman et al., 2022; Lison et al., 2024) as many countries and health authorities faced similar challenges. The present work builds upon a systematic comparison of nowcasting methods in a real-time application to German COVID-19 hospitalization incidences (Wolffram et al., 2023). For this study, a complete set of daily probabilistic nowcasts from eight models and over a six-month period (from November 2021 to April 2022) was compiled, which we use to study two related research questions.

Firstly, we develop statistical post-processing methods for infectious disease nowcasts, similar to existing methods from weather forecasting (Gneiting et al., 2005; Schulz et al., 2021). Post-processing aims at correcting systematic shortcomings of predictions from individual models, like biases and dispersion errors. In our case study, underdispersion of forecasts, i.e., too narrow prediction intervals, was the most common shortcoming of models. In order to suitably transform model outputs, an additional statistical model is fitted to past nowcast and observation pairs. Secondly, we address ensemble nowcasts, which combine different individual nowcasting models. Simple unweighted nowcast ensembles have been found to perform favourably in Wolffram et al. (2023), raising the question whether further improvements can be achieved by weighting different models in a suitable manner. Data-driven weighting of ensemble members is an active area of research in infectious disease forecasting (Yamana et al., 2016; Reich et al., 2019; Reis et al., 2019). For instance the US CDC have used weighted forecast ensembles to inform public health decision making during the COVID-19 pandemic (Ray et al., 2023). To date, however, evidence on the benefits relative to simple unweighted ensembles remains mixed (Bracher et al., 2021a; Ray et al., 2023). This echoes the broader statistical literature, where it has been pointed out that the estimation of ensemble weights comes at a cost which may not necessarily be outweighed by the benefits (Claeskens et al., 2016).

In our application to German COVID-19 hospitalization incidences, we find that post-processing of infectious disease nowcasts leads to quite consistent improvements across nowcasting methods and horizons. This holds both for nowcast calibration in terms of interval coverage rates and for score-based evaluation. Data-driven weighting of nowcast ensembles, on the other hand, proves to be a very challenging task. Exploring a variety of weighting methods, we find consistent improvements in calibration. In terms of evaluation scores, however, we obtain modest improvements for some approaches, and considerable deterioration of performance for others. The more successful weighting schemes tend to be simple, while added complexity rarely translates to improvements.

The remainder of this paper is structured as follows. In Section 2, we describe our applied setting and highlight the challenges of dealing with incomplete data. In Section 3, we introduce the notation used throughout the paper, present the post-processing and ensemble modeling approaches, and discuss the specific challenges posed by data revisions. Section 4 shows the obtained results based on the previously introduced post-processing and ensemble methods applied to the German COVID-19 hospitalization data. Lastly, in Section 5, we discuss our methods and results and comment on the limitations and possible extensions of our work.

2 Motivation: COVID-19 hospitalizations in Germany

For illustration we briefly sketch our applied nowcasting setting, to which we will return in Section 4. We are concerned with the *7-day COVID-19 hospitalization incidence* (German Federal Ministry of Health, 2023), which played an important role in pandemic planning in Germany especially in fall and winter 2021/2022. Temporarily, this indicator even served to determine the necessary level of non-pharmaceutical interventions via a set of thresholds (German Federal Government, 2023). The 7-day hospitalization incidence is defined as the number of new COVID-19 cases registered by local health authorities over a 7-day period which led to a hospitalization. Hospital admission is not required to have taken place during the same 7-day period and may in fact occur considerably later. This somewhat unintuitive definition, which was chosen as "a compromise between timeliness and data quality" (Norddeutscher Rundfunk, 2023), implies that hospitalization counts are not aggregated by the day of admission, but by the day of case registration (see Section 2.1 of Wolffram et al. 2023 for a more detailed account). As a consequence, the delay problem described in Section 1 is particularly pronounced for this indicator: an additional delay between the date of case registration and the date of admission is added on top of the actual reporting delay for the hospitalization. This results in strongly incomplete values of the hospitalization incidence for recent dates, and a characteristic dip at the end of the time series. As detailed in Wolffram et al. (2023), data are corrected upwards over quite prolonged periods of time, and may still change months after initial reporting.

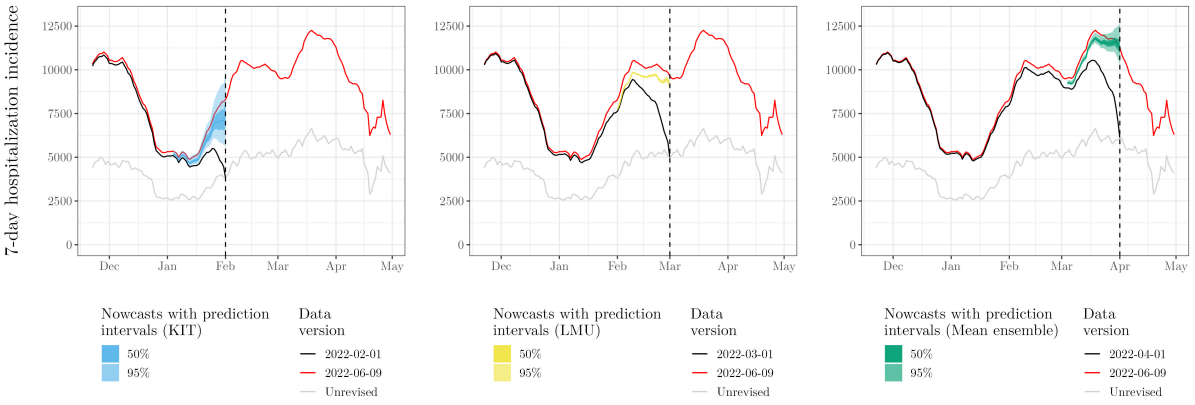

Figure 1: Illustration of the nowcasting task and nowcasts from three different models (`KIT`, `LMU`, and a mean ensemble) on February 01, March 01, and April 01, 2022, respectively. Black lines show data as available in real time on the respective forecast date, with the characteristic dip due to delays. Red line shows the data as completed later (40 days after the end of the displayed period). Point nowcasts and 50% and 95% uncertainty intervals are shown in colors.

Figure 1 illustrates the nowcasting task and nowcasts generated in real time using three different methods. The black lines show data as available when the respective nowcast was issued. The red line shows a later version of the time series including retrospective completions. Light grey lines show unrevised data where for each date only the initial value reported on that same date is shown (implying that the latest value of the black line coincides with the grey line). Nowcasts, i.e., predictions of completed incidence values, are shown as coloured bands. These have been collected in the *German COVID-19 Nowcast Hub* (`https://covid19nowcasthub.de`), a collaborative modelling project involving eight independent modelling groups. The Nowcast Hub aimed to provide reliable assessments of recent trends via daily updated nowcasts, but also to conduct a systematic methods comparison (Wolffram et al., 2023). The analyses in the present paper will be based on the study period of this comparative evaluation

(November 29, 2021, through April 29, 2022). Overall, we consider eight different individual (i.e., stand-alone) models from this project, which are described briefly in Appendix A. Moreover, unweighted median and mean ensembles are available, see Section 3.5.1.

As can be seen from Figure 1, different methods produce nowcasts with different characteristics. The `KIT` model, shown in the left panel, issued rather wide uncertainty intervals, while the intervals from the `LMU` model (middle panel) were considerably more narrow. The right panel shows the median ensemble nowcast, which represents an unweighted combination of all eight models and has uncertainty intervals of medium width.

3 Methods

In this section, we introduce basic concepts and notation on probabilistic disease nowcasts and their evaluation. Moreover, we describe the methods employed for post-processing and ensemble forecasting, and discuss the particularities arising from the fact that observations are subject to revisions over prolonged periods of time.

3.1 Notation for probabilistic nowcasting

Denote by $\{x_t\}_{t\in\mathcal{T}}$, where $\mathcal{T} = \{1, \cdots, T\}$, a daily time series of interest. In our application, x_t is a rolling sum over trailing 7-day windows, but is nonetheless indexed by days. We assume that x_t is not directly observable in real time. Instead, on day t, we can only observe a preliminary version x_t^t. This value is subsequently revised each day, with x_t^{t+d} denoting the value as available on day $t + d$. We assume that data are only subject to revisions up to D days after the fact, so that

$$x_t = x_t^{t+D}. \tag{1}$$

In our application, revisions arise from delayed reports of hospitalizations and are thus typically upwards (though in practice downwards corrections occur exceptionally). The hospitalizations added to the record with a delay of d days correspond to the increments $x_t^{t+d} - x_t^{t+d-1}$. It is common to arrange these in a so-called *reporting triangle* – compare e.g., Günther et al. (2021). For our paper, however, we use the above generic notation which facilitates the display of the proposed methods.

At time $t^\star$, the nowcasting task consists in predicting $x_{t^\star}, \ldots, x_{t^\star - D+1}$, i.e., the final values of those data points which are still subject to revisions. Nowcasts are typically based on the corresponding partial data $x_{t^*}^{t^*}, \ldots, x_{t^*-D+1}^{t^*}$, but may also take into account other information available at t^*. Throughout the paper, we will consider probabilistic nowcasts, stored as quantiles at a pre-defined set $\mathcal{A}$ of levels. For $\alpha \in \mathcal{A}$, we denote the predictive α quantile for x_t issued by model m at time $t^\star$ by

$$q_t^{t^\star,\alpha,m}.$$

In the following, we refer to day t^* as the "nowcast date" and day t as the "target date." Moreover, we denote by $h = t - t^\star$ the *horizon* of the nowcast, meaning that a 0-day ahead nowcast refers to day t^*, a -1 day ahead nowcast to the previous day $t^* - 1$ and so on.

3.2 Evaluation metrics

Post-processing and ensemble weighting typically require assessing the historical predictive performance of one or several models. To evaluate nowcasts we will employ the weighted interval score (WIS, Bracher et al. 2021a), which has been widely used to evaluate quantile-based predictions during

the COVID-19 pandemic (e.g., Cramer et al. 2022). Denote by F a predictive distribution issued for a quantity x, and by $\{q^{\alpha_1}, \cdots, q^{\alpha_A}\}$ the quantiles of F at the $A = |\mathcal{A}|$ pre-defined levels. The WIS is built upon the standard piece-wise linear quantile score (Gneiting and Raftery, 2007), also referred to as the pinball score. For a single quantile level α, it is given by

$$\mathrm{QS}_\alpha(q^\alpha, x) = 2 \cdot [\mathbb{1}(x \leq q^\alpha) - \alpha] \cdot (q^\alpha - x), \tag{2}$$

where $\mathbb{1}$ denotes the indicator function. The WIS is then defined as the average quantile score achieved across quantile levels,

$$\mathrm{WIS}(q^{\alpha_1}, \ldots, q^{\alpha_A}, x) = \frac{1}{A} \sum_{a=1}^{A} \mathrm{QS}_{\alpha_a}(q^{\alpha_a}, x_t). \tag{3}$$

The WIS is negatively oriented, meaning that smaller values are better. It represents a quantile-based approximation of the commonly used continuous ranked probability score (CRPS; Gneiting and Raftery 2007) and can be interpreted as a probabilistic generalization of the absolute error. It is a proper scoring rule, meaning that it encourages honesty of forecasters. As detailed in Bracher et al. (2021a) and described in Appendix B, the WIS can be split into components for forecast spread, overprediction, and underprediction. This decomposition will be used to characterize biases and dispersion errors of different models.

3.3 Including preliminary observations in nowcast evaluations

In disease nowcasting, information on the target quantity accumulates in a more gradual manner than in classical forecasting settings. On day t^*, the weighted interval score (or any other evaluation score) cannot be evaluated for target dates $t^* - 1, \ldots, t^* - D$, even though some new information on the values $x_{t^*-1}, \ldots, x_{t^*-D+1}$ has already accumulated, with e.g., x_{t^*-D+1} usually almost exactly known. Simply ignoring the respective nowcasts in the assessment of past performance would mean giving up on this information, which due to its recency may be particularly relevant. We will asses the two following approaches to integrate it into our post-processing or ensemble weighting methods.

- **Simple imputation:** In order to complete the partial observations $x^{t^*}_{t^*-1}, \ldots, x^{t^*}_{t^*-D+1}$, using the information on available on day t^* an obvious strategy is to use up-to-date nowcasts. We thus replace the values in question by pseudo observations defined as

$$\tilde{x}^{t^*}_t = q^{t^*, 0.5, \texttt{mean}}_t,$$

  i.e., current predictive medians. We use nowcasts from the unweighted mean ensemble, denoted by `mean` which we know has rather reliable performance (Wolffram et al., 2023). Intuitively speaking, rather than comparing nowcasts issued during the last $D - 1$ days to the truth, we assess how strongly they already had to be revised in light of new data.
- **Imputation with uncertainty:** The simple imputation approach neglects the uncertainty remaining in the most recent unweighted ensemble nowcasts. In a second, more sophisticated approach, we compare past nowcasts to all nowcast quantiles $q^{t^*, \alpha_1, \texttt{mean}}_t, \ldots, q^{t^*, \alpha_A, \texttt{mean}}_t$ issued on day t^*. This can be done using a generalization of the weighted interval score described in Appendix C. It is inspired by a similar generalization of the CRPS which has been suggested by Friederichs and Thorarinsdottir (2012) to account for observation errors in meteorological forecast evaluation.

3.4 Post-processing individual models

We now address the improvement of nowcasts from individual models via statistical post-processing. To this end, we employ a simple re-scaling approach. Specifically, at nowcast time t^*, the predictive α quantile issued by a given model for target time t is transformed as

$$q_t^{t^*,\alpha,\text{post}} = x_t^{t^*} \quad + \quad \phi^{t^*-t,\alpha} \times \left(q_t^{t^*,\alpha} - x_t^{t^*}\right), \tag{4}$$

where we suppressed the index m for the model to simplify notation. Scaling is thus only applied to the difference between the currently known number of hospitalizations $x_t^{t^*}$ and the predicted final value $q_t^{t^*,\alpha}$. In our application we will constrain $\phi^{h,\alpha} > 0$, which ensures that the nowcast quantile cannot fall below the already known number of hospitalizations. In the most general formulation, the scale parameter $\phi^{h,\alpha}$ is specific both to the quantile level α and the nowcast horizon h. While we also consider a more parsimonious formulation where a parameter ϕ^{α} is shared across horizons $h = 0, \ldots, -D+1$, we always keep it specific to the quantile level α. This is done because the correction of dispersion errors, which is a central aim of post-processing, requires upward scaling of some quantile levels and downward scaling of others.

The value of $\phi^{h,\alpha}$ is determined via score minimization over a training period $\mathcal{R}$, i.e., it is chosen such that the objective

$$\sum_{r \in \mathcal{R}} \text{QS}_\alpha(x_r, q_r^{r+h,\alpha,\text{post}}) \tag{5}$$

is minimized. The training period $\mathcal{R}$ includes days $t^* - R, \ldots, t^* - D$ for which definitive observations are already available. In practice we set R to the minimum of the number of days for which individual-model nowcasts are already available on day t^* and some maximum number (e.g., $R = 90$). Depending on the strategy chosen to handle incomplete data, $\mathcal{R}$ may in addition contain days $t^* - D + 1, \ldots, t^* - 1$, for which pseudo-observations need to be employed in the evaluation. As detailed in Section 3.3, these may be given either by single numerical values or a set of predictive quantiles. In the latter case, the respective summands in Equation (5) are given by an adapted version of the quantile score described in Appendix C. Regarding the minimization of the target function, we follow Ray et al. (2023) in using a grid search approach to determine $\phi^{h,\alpha}$.

3.5 Combination of nowcasting models

To combine nowcasts from M different models into an ensemble we will use mappings of the general form

$$q_t^{t^\star,\alpha,\text{ens}} = f(q_t^{t^\star,\alpha,1}, \ldots, q_t^{t^\star,\alpha,M}),$$

i.e., the ensemble quantile is computed from the respective member quantiles at the same level. In the following sections, we elaborate on different specifications of f, from simple unweighted schemes to more sophisticated weighted and data-driven schemes. As discussed e.g., Ray et al. (2023), the space of possible formulations and parameterizations is vast, and complexity can be added in many different ways. Our rationale is to explore a set of distinct, but reasonably simple approaches which could be operated in practice.

3.5.1 Unweighted combination

The simplest approach is given by unweighted ensembles, as used in Wolffram et al. (2023). For the mean ensemble, predictive quantiles from models $m = 1, \ldots, M$ are aggregated as

$$q_t^{t^*,\alpha,\text{ens}} = \frac{1}{M} \sum_{m=1}^{M} q_t^{t^*,\alpha,m}. \tag{6}$$

As in Wolffram et al. (2023), we will also consider a median ensemble, which uses the median rather than the mean to aggregate quantiles from different models.

3.5.2 Post-processing-based approaches

An obvious approach to improve upon the unweighted ensemble is to harness the post-processing methods described in Section 3.4. As the order of post-processing and combination of forecasts is not interchangeable, we consider two approaches:

- **Post-process, then combine:** If post-processing can improve upon each model individually, one may expect a combination of post-processed models to be superior. We thus consider an unweighted mean and median ensembles of the post-processed members.
- **Combine, then post-process:** Alternatively, the different models can be combined to an unweighted mean or median ensemble first, which is subsequently subject to post-processing. This approach is computationally cheaper as post-processing only needs to be run once.

3.5.3 Direct inverse-score weighting

A second rather straightforward strategy consists in "direct inverse-score weighting" (DISW). We here generalize Equation (6) to

$$q_t^{t^*,\alpha,\text{ens}} = \sum_{m=1}^{M} w^{t^*-t,\alpha,m} \times q_t^{t^*,\alpha,m}$$

while choosing the weights in a heuristic manner, setting

$$w^{h,\alpha,m} = \frac{\frac{1}{\overline{\text{QS}}_{\mathcal{R}}^{h,\alpha,m}}}{\sum_{i=1}^{M} \frac{1}{\overline{\text{QS}}_{\mathcal{R}}^{h,\alpha,i}}}. \tag{7}$$

Here, $\overline{\text{QS}}_{\mathcal{R}}^{h,\alpha,m}$ is the average quantile score for model m, quantile level α and horizon h days computed on a training period $\mathcal{R}$ as defined in Equation (5). The rationale is that models with good historical performance (low average scores) should receive larger weights. As in Section 3.4, we will assess both a version with weights specific to horizons and a simplified version where shared weights $w^{m,\alpha}$ are employed. Inverse-score weighting has been used for COVID-19 forecasts in Bracher et al. (2021b), where in turn it had been borrowed from the meteorological literature (Zamo et al., 2021). An advantage of this approach is that it does not require any costly optimization.

3.5.4 Adjustable inverse-score weighting

Direct inverse score weighting has two obvious limitations. Firstly, it makes a strong assumption on how weights should depend on past WIS scores. Secondly, as it is a convex combination of the models, no correction for biases shared by all members is possible. If, for instance, all member models show a downward bias, then so will the ensemble. We therefore render the approach more flexible by

introducing two additional parameters. We will refer to this as "adjustable inverse-score weighting" (AISW). Combining ideas from Equations (4) and (7), we set

$$q_t^{t^*,\alpha,\text{ens}} = x_t^{t^*} + \phi^{t^*-t,\alpha} \times \sum_{m=1}^{M} w^{t^*-t,\alpha,m} \times (q_t^{t^*,\alpha,m} - x_t^{t^*}) \tag{8}$$

with weights defined as

$$w^{h,\alpha,m} = \frac{\left(\frac{1}{\overline{\text{QS}}_{\mathcal{R}}^{h,\alpha,m}}\right)^{\theta^{h,\alpha}}}{\sum_{i=1}^{M}\left(\frac{1}{\overline{\text{QS}}_{\mathcal{R}}^{h,\alpha,i}}\right)^{\theta^{h,\alpha}}}. \tag{9}$$

Here, $\phi^{t^*-t,\alpha}$ can shift predictive quantiles up and down. As in the post-processing scheme from Section 3.4, scaling is only applied to the predictions of yet-to-observe hospitalizations, while the current count $x_t^{t^*}$ is not modified. If only $M = 1$ model is available, the approach is thus equivalent to (4). The parameter $\theta^{h,\alpha}$ steers how strongly weights depend on differences in past scores. A value of 0 implies equal weighting as in (6), while positive values assign more weights to models with good historical performance. For $\theta^{h,\alpha} = 1$, the weights correspond to the DISW approach (7). We also allow $\theta^{t^*-t,\alpha}$ to be negative, in which case models with worse past scores would receive more weight. As before, we will also apply a simplified parameterization where weights are shared across nowcast horizons. The weights and scaling parameter are again determined using a training set of past nowcasts and observations as denoted in Equation (5).

The described approach is a variation of the one used in Ray et al. (2023). It serves to keep the number of parameters moderate and circumvent identifiability problems arising from the typically strong correlations between predictive quantiles from different models (indeed, unconstrained quantile regression was poorly behaved in our application). We changed the transformation of average scores from an exponential to a power relationship as we wanted Equation (7) to nest into the more general case. In exploratory analyses we found that the two approaches behave rather similarly in practice.

3.5.5 Top-n model selection

An alternative to explicit weighting of models is to restrict the ensemble to a pre-specified number n of models which historically have shown the best performance. At time t^* and for each quantile level α and horizon h, we thus order models according to the average quantile score $\overline{\text{QS}}_{\mathcal{R}}^{h,\alpha,m}$. Then, the n best-performing models are retained and averaged into a mean or median ensemble without further weighting. As for the other approaches, we will again consider a simplified version of the approach where all horizons are treated jointly.

4 Application to German COVID-19 hospitalizations

In the following we provide details on the COVID-19 hospitalization nowcasting task from Section 2 and highlight differences to previous work. This is followed by a performance assessment for the various post-processing and combination methods. To keep the presentation structured despite the large number of considered approaches, we provide some interpretation of the results already in the respective subsections rather than the discussion section.

4.1 Technical description of the nowcasting task

Nowcasting targets and horizons. Paralleling Wolffram et al. (2023), we will consider nowcasts at horizons $h = 0, \ldots, -28$ days. These are available at the national level, for the 16 German states and for 7 age groups. In Wolffram et al. (2023), the target (i.e., ground truth used in evaluation) for nowcasts referring to target date t was pre-registered as the corresponding value of the 7-day hospitalization incidence according to a data set published considerably later. More specifically, data of 8 August 2022, i.e., 100 days after the end of the last nowcasting date were used. This was based on the assumption that all incidence values would have stabilized after 100 days. In reality, however, the data kept being revised upwards – compare Section 3.7 in Wolffram et al. (2023). This is disadvantageous for several reasons. From a public health standpoint, it is unclear if these very late reports are meaningful for real-time situation assessment, see discussion in Wolffram et al. (2023). From a statistical standpoint, it is undesirable that the evaluation of nowcasts depends on which data version is used as the ground truth. Moreover, the data could be corrected for a longer time for target dates early in the study period compared to later ones. In Wolffram et al. (2023) it was therefore concluded that a more meaningful definition of the nowcasting task would have been as in Equation (1), i.e., as the number of hospitalizations reported up to a maximum delay of D days. In the following, we will use this definition with $D = 40$ as this was the maximum delay the modelling teams assumed in their statistical analysis (for the `ILM` team, who assumed a larger value of $D = 84$, we obtained adjusted nowcasts with a matching maximum delay). While it may seem odd to adjust the nowcasting target retrospectively, we argue that this is meaningful for the present work. Unlike in Wolffram et al. (2023), our goal is not to assess real-time performance for a pre-defined task. Rather, we aim for a meaningful test bed for methods development. We consider the re-defined target helpful as it is technically more sound and better aligned with what the various models were set up to achieve in practice.

Study period. We consider nowcasts generated in a daily rhythm from November 29, 2021, to April 29, 2022. As all data-driven post-processing and ensembling methods require some historical pairs of nowcasts and observations for training, we hold out the first 70 days of this period. The performance evaluation is conducted over the remaining time period (February 8, 2022 through April 29, 2022; i.e., 81 days). By leaving out 70 days, we ensure that a minimum of 30 days of complete data is available for training the post-processing and ensembling methods (even after excluding the most recent $D = 40$ days with yet incomplete observations).

4.2 Performance of original nowcasts from Wolffram et al. (2023)

We start by summarizing the performance of the eight individual models and two ensembles from Wolffram et al. (2023) in our adapted setting. Figure 2 shows nowcasts issued by different models over time for two horizons (0 and −14 days). Figure 3 displays average WIS values and interval coverage fractions for national-level and stratified nowcasts. Note that the `ILM` and `RKI` teams did not report nowcasts for states and ages groups, respectively. This figure is similar to Figure 13 from Wolffram et al. (2023), but, as we left out the nowcasts from the first 70 days for training purposes, refers to a shorter evaluation period.

While the results are discussed in detail in Wolffram et al. (2023), we point out some relevant aspects. Most individual models have small spread components of the WIS, indicating narrow prediction intervals (left column). The is in line with the corresponding interval coverage fractions, which are considerably below nominal levels (right column). This pattern is particularly pronounced for the `LMU`, `RIVM` and `RKI` models, while the `KIT` model is somewhat better calibrated (see also the respective panel ins Figure 2). The `SZ` model has a large underprediction component of the average WIS, suggesting a downward bias. This holds especially for more distant horizons, as illustrated by the −14 day ahead

nowcasts in Figure 2. These overall patterns are shared across the national level and stratified nowcasts. We note that the WIS values for the stratified targets are lower on average because the WIS is scale-dependent.

The unweighted mean and median ensembles achieve substantially better performance than all individual models in terms of average WIS. Also, their prediction intervals, while not reaching nominal coverage, are better calibrated. The difference between the median and mean formulation are small. These patterns hold both at the national level and for nowcasts stratified by age or state.

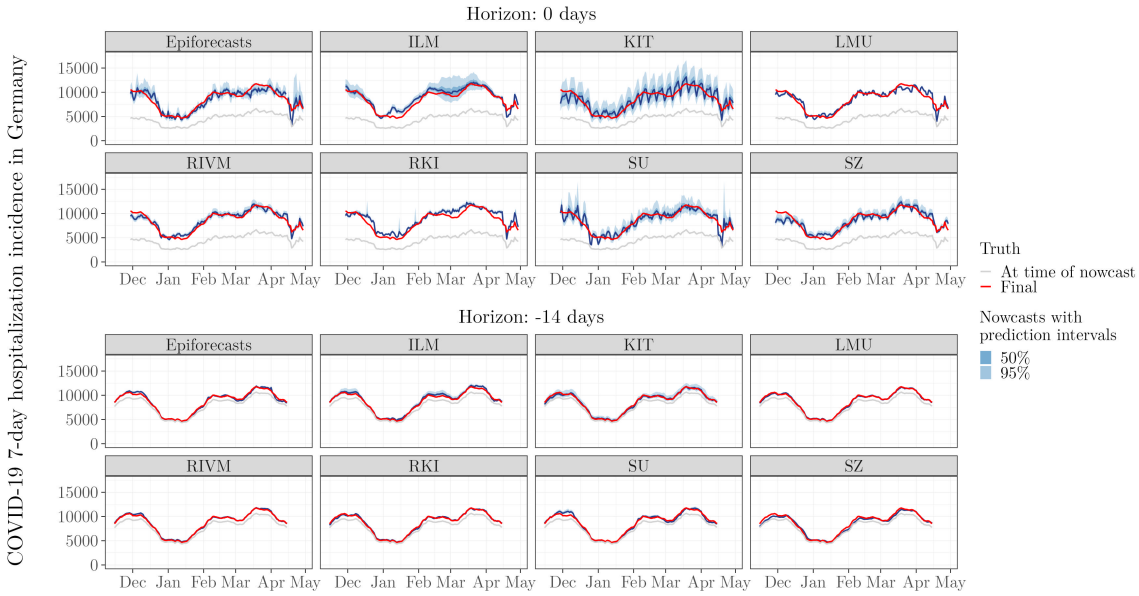

Figure 2: National-level nowcasts (blue) with horizons of 0 and −14 days for the eight individual models, by target date. The red line shows the nowcasting target, i.e., the number of COVID-19 7-day hospitalization cases after 40 days of retrospective corrections. The grey lines show the reported incidence counts at the time of nowcasting, i.e.,after 0 (top) and 14 days (bottom), of retrospective corrections. Blue shaded areas represent nowcast intervals. This figure parallels Figures 5 and 6 from Wolffram et al. (2023).

4.3 Performance of post-processed individual models

Based on the methodology described in Section 3.4, we post-processed the nowcasts from all eight individual models. In our main analysis, we used a maximum value of $R = 90$ days for training, i.e. we only used fairly recent nowcasts and observation pairs. In the Supplementary Material, we present results for a maximum of $R = 60$ days and without any maximum value for R (finding that the improvements in average WIS when using data from more than $R = 90$ days are negligible; see Figure SF1). Varying the analytical options described in Section 3.3, we investigated the post-processing approach with a total of four different settings (see upper part of Table 1). These differ in whether and how yet incomplete observations are included into the training set (see Section 3.3) and whether the scaling parameters are shared or differ across horizons. For each version, we introduce a label which we will use for referencing in the following (set in `typewriter` font).

The average WIS and empirical coverage proportions for the eight individual models and four post-processing variations at the national level are presented in Figure 4 (`PP4`) and Supplementary Figures SF2–SF4. Comparing Supplementary Figures SF2 (`PP1`) and SF3 (`PP2`), we see that including yet incomplete observations into the training set is beneficial, yielding improved WIS performance for

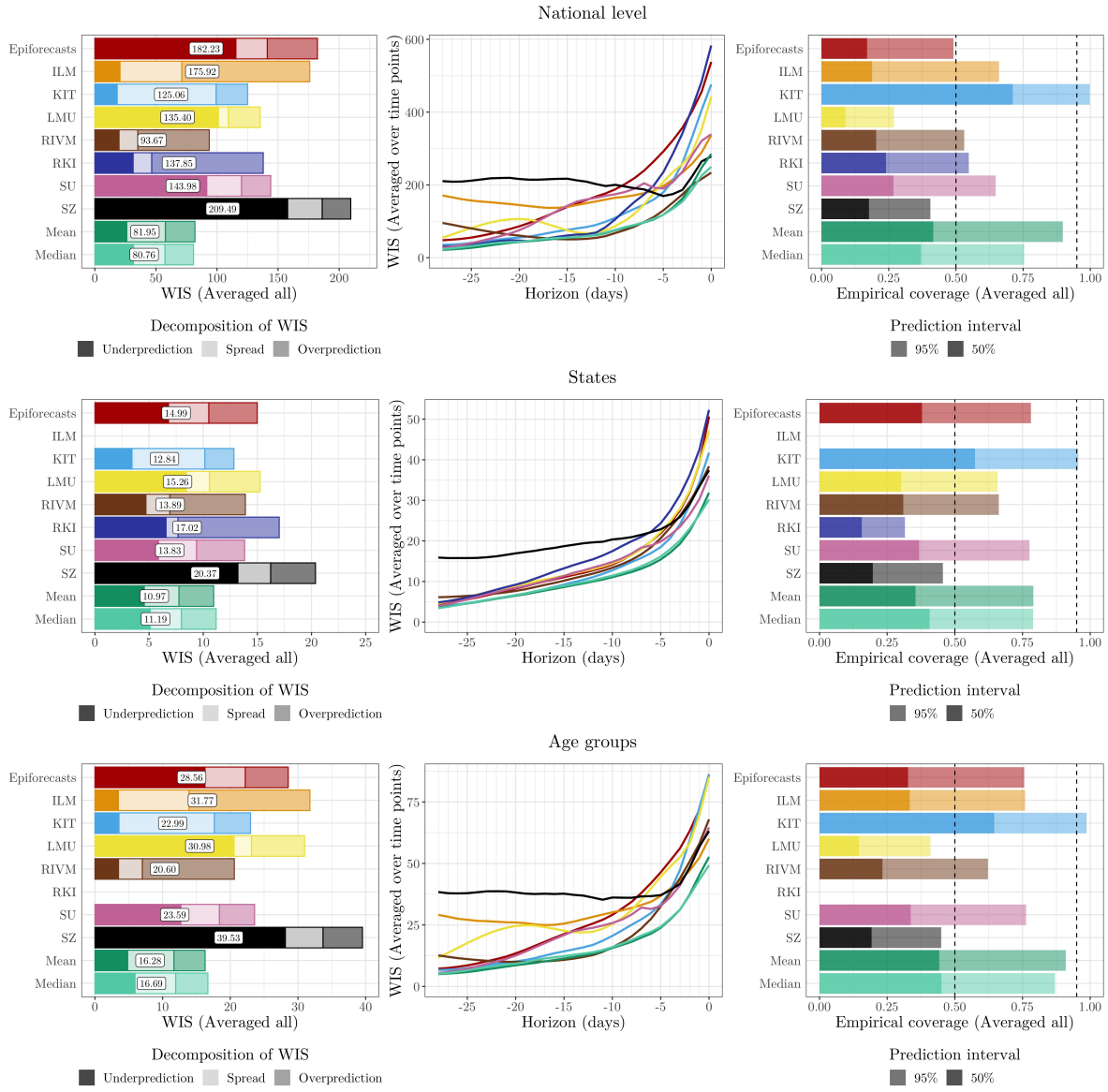

Figure 3: Model performance of original models and ensembles from Wolffram et al. (2023). Left: WIS (averaged over time points and horizons), split into components for underprediction, spread, and overprediction. Middle: WIS by nowcast horizon (averaged over time points). Right: Empirical coverage proportions (averaged over time points and horizons). The results are reported for the national level (top row) and averaged across states (middle row) and age groups (bottom row).

almost all models. The more sophisticated imputation with uncertainty (`PP3`, Figure SF4) considerably increased computation times, but compared to simple imputation (`PP2`) had very limited impact on the nowcasts and their performance. The more flexible version `PP4` with separate handling of different horizons (Figure 4) results in slightly better overall performance. The following description of results is therefore focused on this version but the qualitative results also apply to the three others.

After post-processing, the average WIS values of all models decrease, the WIS components are more balanced and the empirical coverage rates are closer to the nominal values. We discuss these aspects in detail for the `LMU` and `SZ` models which, as mentioned in Section 4.2, have specific dispersion errors and biases. For `LMU`, we notice that the spread component of the WIS is larger than before, implying wider prediction intervals. This is the intended effect of post-processing, and also observed for the `RIVM`

Table 1: Post-processing and combination approaches assessed in Section 4. All methods are fitted to national-level data, methods marked with a star symbol ($\star$) are moreover applied to stratified data (age groups and states). The "Label" column contains a short identifier used for brevity in the remaining text and figures.

Method	Sec.	Label	Settings
Post-processing			
Re-scaling	3.4	`PP1`	Scaling parameter ϕ^α shared across horizons while discarding incomplete observations
		`PP2`	Scaling parameter ϕ^α shared across horizons with simple imputation
		`PP3`	Scaling parameter ϕ^α shared across horizons with imputation with uncertainty
		`PP4`	Scaling parameter $\phi^{t^*-t,\alpha}$ varying over horizons with simple imputation
Combination			
Unweighted	3.5.1	`Mean`	Mean ensemble*
		`Median`	Median ensemble*
Post-processing-based	3.5.2	`Post-Mean`	Mean ensemble of post-processed models (`PP4`)
		`Post-Median`	Median ensemble of post-processed models (`PP4`)
		`Mean-Post`	Post-processed (`PP4`) mean ensemble
		`Median-Post`	Post-processed (`PP4`) median ensemble
DISW	3.5.3	`DISW1`	Weights $w_t^{\alpha,m}$ shared across horizons, discarding incomplete observations
		`DISW2`	Weights $w_t^{\alpha,m}$ shared across horizons, simple imputation
		`DISW3`	Weights $w_t^{\alpha,m}$ shared across horizons, imputation with uncertainty*
		`DISW4`	Weights $w_t^{h,\alpha,m}$ varying over horizons, simple imputation*
AISW	3.5.4	`AISW1`	Weights $w_t^{\alpha,m}$ and scaling parameter ϕ^α shared across horizons, discarding incomplete obs.
		`AISW2`	Weights $w_t^{\alpha,m}$ and scaling parameter ϕ^α shared across horizons, simple imputation
		`AISW3`	Weights $w_t^{\alpha,m}$ and scaling parameter ϕ^α shared across horizons, imputation with uncertainty*
		`AISW4`	Weights $w_t^{h,\alpha,m}$ and scaling parameter $\phi^{h,\alpha}$ varying over horizons, simple imputation*
Select-n	3.5.5	`Select-n-Mean1`	Mean ensemble, model selection shared across horizons, simple imputation
		`Select-n-Median1`	Median ensemble, model selection shared across horizons, simple imputation
		`Select-n-Mean2`	Mean ensemble, model selection independent for horizons, simple imputation
		`Select-n-Median2`	Median ensemble, model selection independent for horizons, simple imputation

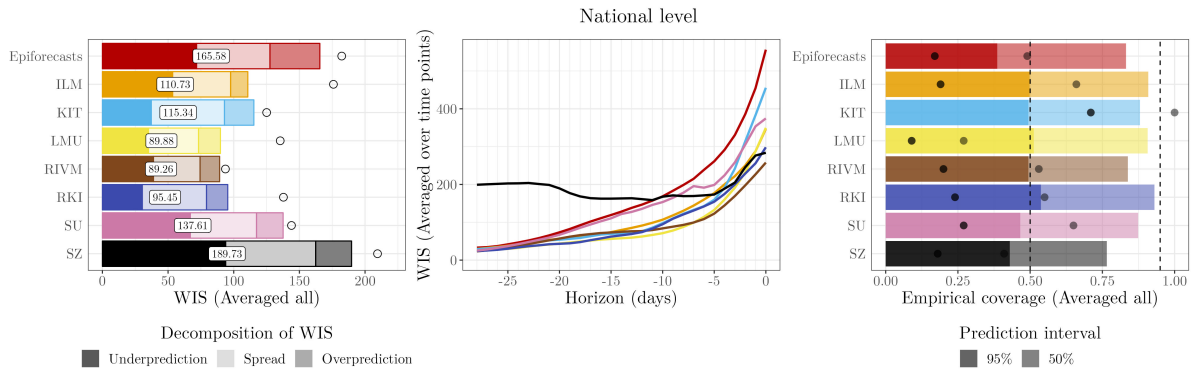

Figure 4: Performance of post-processed (`PP4`) individual-model nowcasts compared to the original versions, national level. Left: WIS (averaged over time points and horizons) split into components for underprediction, spread and overprediction. Middle: WIS by nowcast horizon (averaged over time points). Right: Empirical coverage proportions (averaged over time points and horizons). In the left and right panel, circles (○) represent the results for the original models before post-processing, i.e., as in Figure 3.

and `RKI` models. We illustrate the widening of nowcast intervals for same-day nowcasts with $h = 0$ in Figure 5 (first row, left column; consider the respective panel of Figure 2 for comparison). The score improvements are very consistent over nowcast horizons and dates (Figure 5, first row, middle and right columns). For the `SZ` model, although the overall WIS is not drastically improved, the underprediction component of the WIS is much smaller and the empirical coverage rates are better than before. As can be seen for $h = -14$ in the second row of Figure 5, the post-processed `SZ` nowcasts no longer display a clear bias. The improvement in average WIS values is pronounced for more distant horizons, while for short horizons post-processing actually leads to a minor deterioration of performance.

For the other models (Supplementary Figures SF5–SF10), there are improvements in average

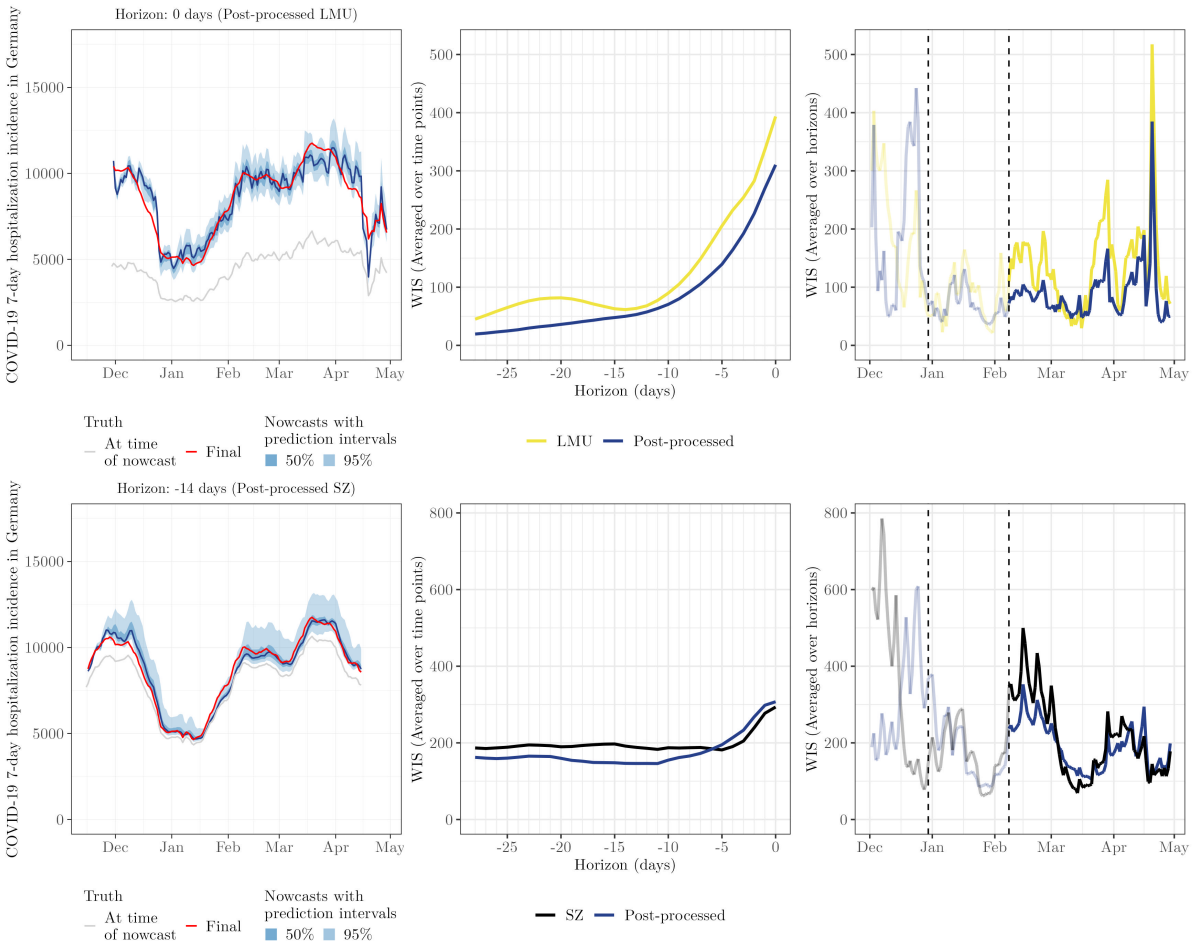

Figure 5: Left column: Same-day nowcasts for the post-processed `LMU` model (first row) and −14-day nowcasts for the post-processed `SZ` model (second row). All nowcasts are at the national level and based on the post-processing scheme `PP4`. Grey lines show the data as available when the nowcasts were issued. Red lines represent the nowcasting target, i.e., the incidence count after 40 days of retrospective corrections. Middle column: Average WIS before and after post-processing, by nowcast horizon. Right column: WIS (averaged over horizons) before and after post-processing, per target date. The two dashed vertical lines represent 30 December, 2021, i.e., the earliest target date, and February 8, 2022, i.e., the first nowcast date of the evaluation period. Scores before 8 February (greyed out) only partly enter into the reported average scores, but are shown for illustration.

WIS, but they are less consistent over time and nowcast horizons. This holds especially for the `KIT` model. As mentioned in Wolffram et al. (2023), the main shortcoming of the `KIT` model is an insufficient handling of weekday patterns, leading to different biases on different days of the week. This aspect cannot be corrected by our simple scaling approach.

4.4 Performance of ensemble approaches

We now turn to the performance of weighted nowcast ensembles. For the various approaches presented in Section 3.5, we again varied the way yet incomplete observations are used and whether parameters are shared across horizons; see the summary in the bottom part of Table 1. Note that due to extensive computing times, only a subset of approaches was applied to the stratified nowcasts (marked by an asterisk*). As before, we used a maximum value of $R = 90$ in the main analyses and assessed

sensitivity to a maximum value of $R = 60$ and no upper limit on R in the Supplementary Material. The performance of the various combination approaches is summarized graphically in Figure 6 for the national level and Figure 7 for age strata and states. A graphical display of nowcasts produced by selected approaches is given in Figure 8. The results for the different approaches are discussed in subsections paralleling the structure of Section 3.5.

4.4.1 Unweighted ensembles

As already evoked in Section 4.2, the unweighted `mean` and `median` ensembles outperform all individual models in terms of average WIS, and most of them in terms of interval coverage. Even after post-processing (Section 4.3), the average WIS of all individual models remains inferior to the unweighted ensembles. For the following, the two unweighted ensembles can thus be seen as the baseline upon which more sophisticated combination approaches should improve.

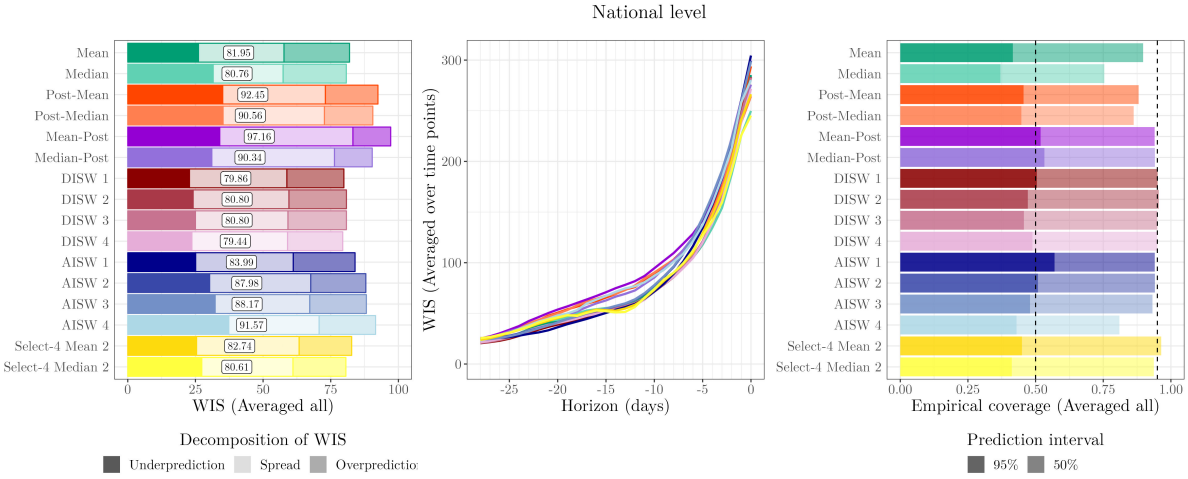

Figure 6: Performance of unweighted and weighted ensemble approaches at the national level. Left: WIS (averaged over time points and horizons), split into components for underprediction, spread and overprediction. Middle: WIS (averaged over time points) by nowcast horizon. Right: Empirical coverage proportions (averaged over time points and horizons).

4.4.2 Post-processing-based approaches

The results achieved by unweighted averaging of post-processed nowcasts (`Post-Mean` and `Post-Median`) and post-processing of unweighted ensembles (`Mean-Post` and `Median-Post`) are quite similar, i.e., the order of post-processing and averaging does not seem to be decisive. In terms of interval coverage, they perform favourably, in particular the post-processed unweighted ensembles. As can be seen from the decomposition of the average WIS in the left panel of Figure 6, this is achieved by a widening of nowcast intervals (see the increased spread components). In terms of average WIS, however, the post-processing-based approaches are not only outperformed by the unweighted ensembles `mean` and `median`, but even some post-processed individual models. This is surprising given that post-processing improved the average performance of all individual models.

While it is hard to provide any definitive explanation for the observed decrease in performance, one possible reason is that post-processing reduces the *diversity* of the ensemble. It is often argued that ensembles work best if their members are diverse and contribute distinct signals (DelSole et al., 2014). By applying the same post-processing scheme to all members, or by glossing over the ensemble

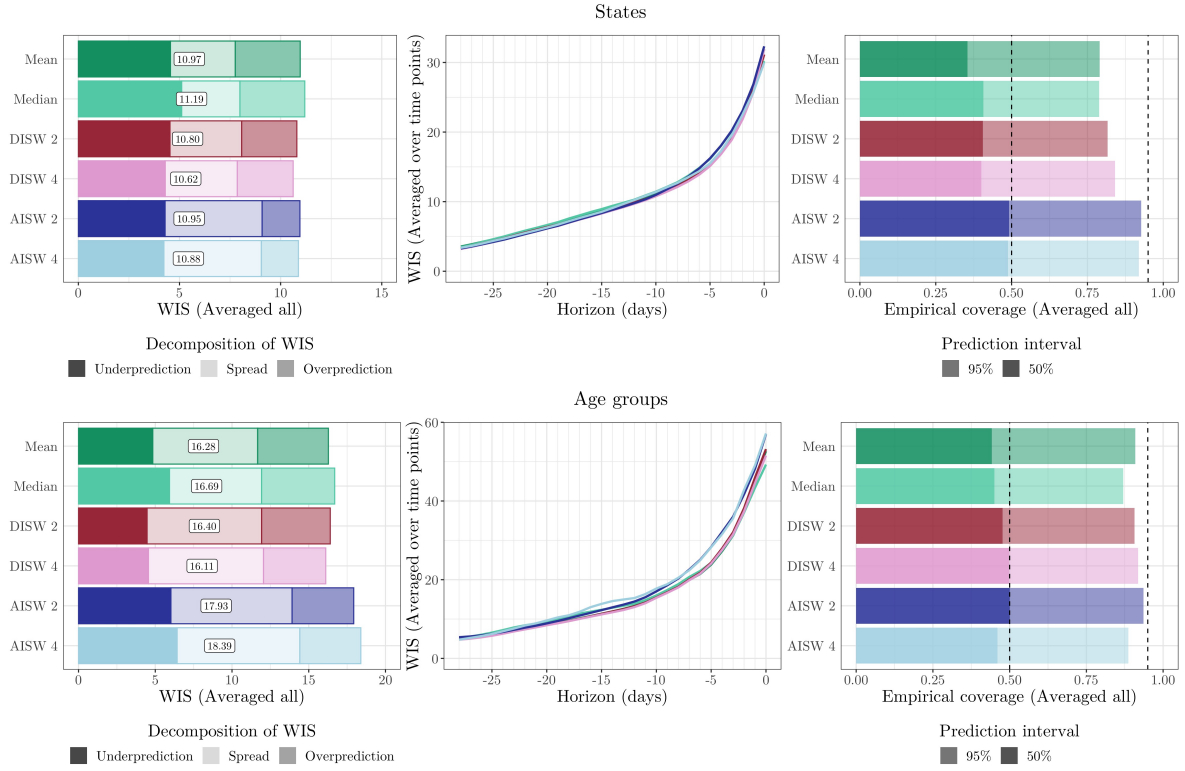

Figure 7: Performance of unweighted and weighted ensemble approaches at the state (top) and age-group levels (bottom; both averaged across strata). Left: WIS (averaged over time points and horizons), split into components for underprediction, spread and overprediction, computed for all stratified ensemble methods. Middle: WIS (averaged over time points) by nowcast horizon. Right: Empirical coverage proportions (averaged over time points and horizons).

nowcast with a single post-processing method, characteristics of the post-processing method may dominate the ensemble characteristics, and diversity may be compromised. In the case of post-processing the unweighted ensembles, it is also possible that the margins for improvement by simple re-scaling are too modest in order to outweigh the cost of estimating scaling factors (we will return to this aspect in Section 4.4.4).

4.4.3 Direct inverse score weighting

The four considered variations of the direct inverse-score weighting overall perform very similarly to the unweighted ensembles, with some modest improvements. The variant `DISW4` (weights varying over horizons, simple imputation) has the lowest average score, but by a margin that should not be interpreted as a meaningful difference. For the nowcasts stratified by age group and state, the results are overall similar, see Figure 7. As we will see in the following, the simple DISW approaches overall achieve the best performance of all considered combination approaches.

The uncertainty intervals of the DISW ensembles are somewhat wider than in the unweighted ensembles; consider again the spread components in the left panel of Figure 6 as well as the illustration of nowcasts in Figure 8. This results in improved calibration at the national and age group levels. Apart from this, however, the DISW forecasts look quite similar to the unweighted `mean` nowcasts.

The weights assigned to the different models are quite close to uniform for the predictive median, see the middle panel of Figure 9. For the 0.025 and 0.975 quantiles, weights are more imbalanced and vary over time. The `RIVM` model, which tends to over-predict (see WIS decomposition in Figure 3), receives

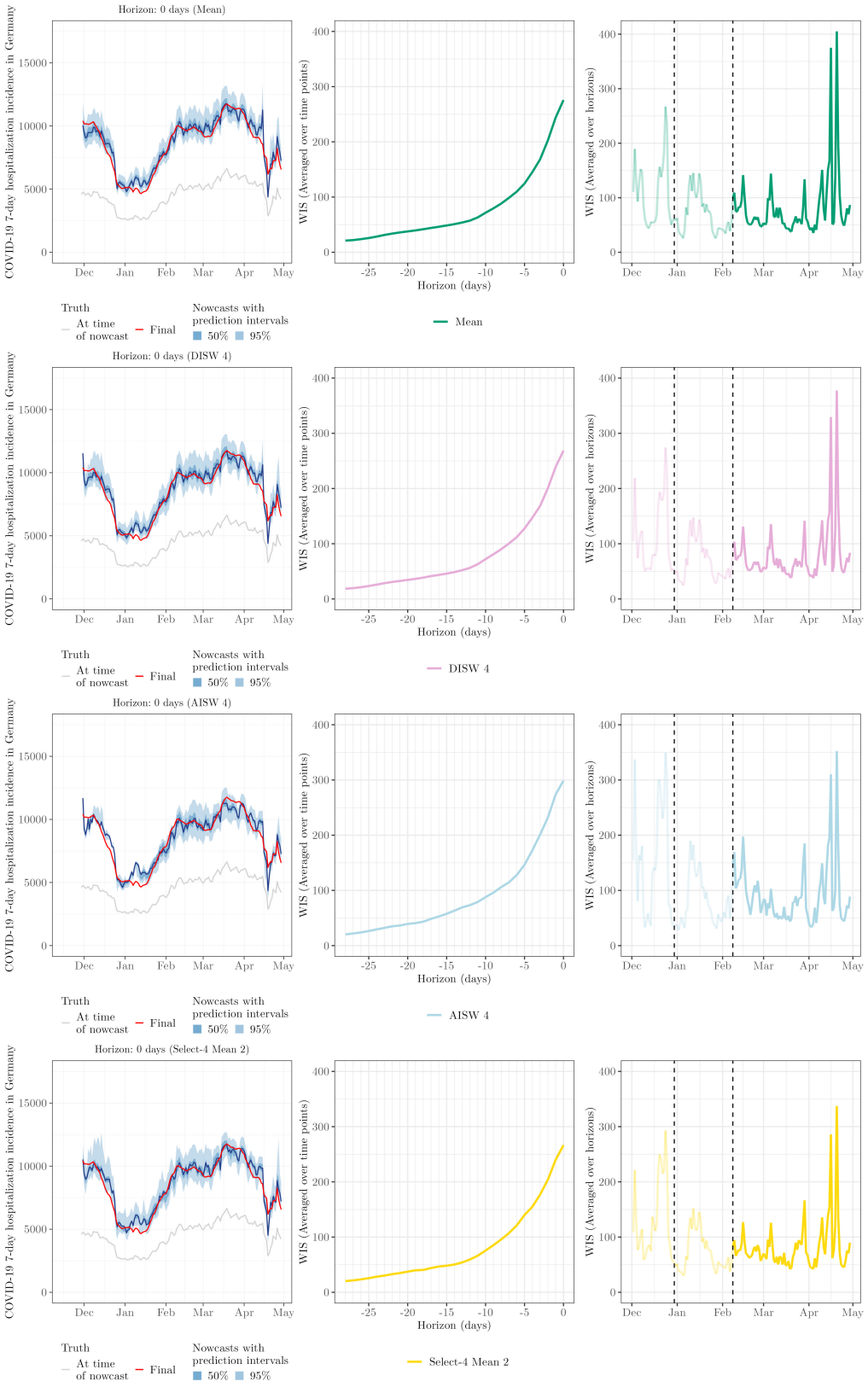

Figure 8: Illustration of same-day nowcasts for the `Mean`, `DISW4`, `AISW4` and `Select-4-Mean2` ensembles. See caption of Figure 5 for details on plot elements and Table 1 for details on the methods specifications.

little weight for the 0.025 quantile. The `LMU` model, on the other hand, receives little weight for the 0.975

quantile, as it tends to underpredict. This explains the aforementioned widening of prediction intervals. To illustrate the behaviour when weights are only based on few historical nowcasts and observations, we also display the initial period 29 November, 2021 through 7 February 2023 (greyed out), which is excluded from the evaluation. As could be expected, the resulting weights fluctuate more strongly during this period. This pattern is more pronounced for the extreme quantiles than the median.

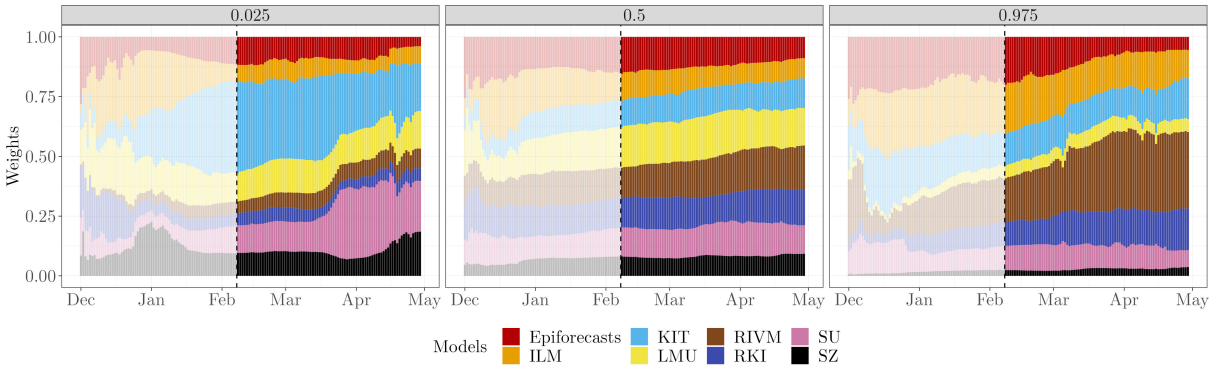

Figure 9: Estimated weights for the 0.025, 0.5, and 0.975 quantiles based on the direct inverse-score weighting method `DISW2` (weights shared across horizons with simple imputation). Weights are shown for the national level. Similarly to Figure 5, results for the period preceding the actual evaluation period are greyed out.

4.4.4 Adjustable inverse score weighting

We now turn to the AISW method, which unlike the DISW approach requires determining scaling and weighting parameters based on past pairs of nowcasts and observations. In practice this resulted in considerably increased computational effort, but did not translate to gains in performance in terms of average WIS. While the difference to the unweighted and DISW ensembles is not drastic, it is consistent across specifications 1 through 4. The interval coverage rates are similar to those of DISW.

Figure 10 shows the estimated weights for setting `AISW2`. The corresponding plots for the other AISW settings, along with the estimated weights aggregated by horizon or quantiles (where applicable), are presented in the Supplementary Material. Several observations can be made from Figure 10. Firstly, the weights are less smooth over time than in Figure 9. In some instances, e.g., in early March for the 0.025 quantile, there are small jumps, which may indicate the presence of several local optima in the objective function (note that our grid search ensures that we do not end up in a local optimum, but the global optimum can "jump" to a different local optimum from one day to the other). For the 0.025 quantiles, the effective model weights (i.e., $\phi^\alpha w_t^{0.025,m}$) sum up to a value below one. The scaling parameter ϕ^α is thus below one and leads to lower (more conservative) ensemble quantiles. For the predictive median, almost no re-scaling takes place, while for the 0.975 quantile there is likewise some downscaling. Compared to Figure 9, the differences between weights received by different models are excerberated, i.e., the AISW ensemble emphasizes models with better historical WIS values even more (meaning that the $\theta^{h,w}$ exceed one). This is especially pronounced for the 0.975 quantile, where the RIVM model receives a large weight towards the end of our study period.

For nowcasts stratified by states and age groups (Figure 7), the performance of the AISW approach is somewhat more favourable. For state-level nowcasts, in which case 16 times more data are available to determine the weights in a data-driven way, the AISW achieves minimally better scores than the unweighted ensemble and minimally worse than the DISW. For age groups, in which case 6 times more data are available to estimate weights, the AISW ensembles again fall behind both the unweighted

and DISW variations.

The results at the national and stratified levels indicate that the estimation of weighting parameters may come at the cost of fluctuating and somewhat instable ensemble weights. The fluctuating nature of the weights may either indicate that there is not enough data to estimate it reliably, or that there is not actually a temporally stable "right" configuration of weights.

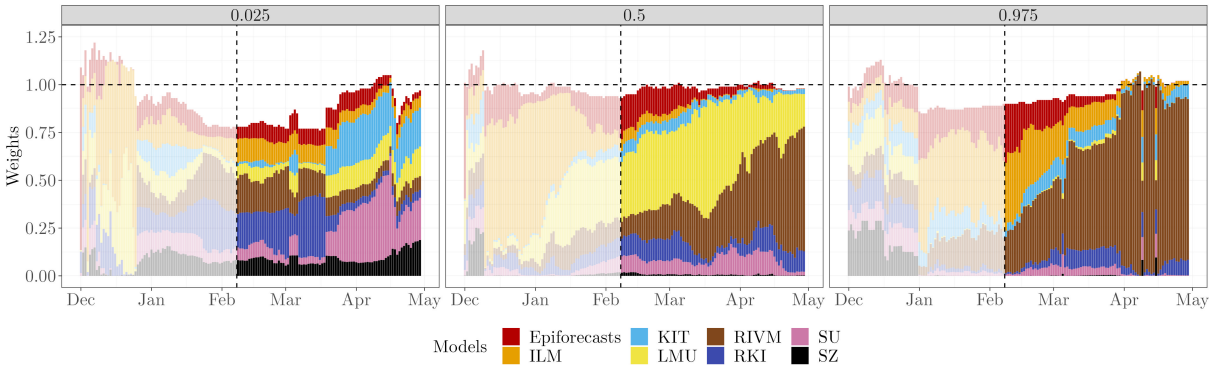

Figure 10: Estimated weights for the 2.5^{th}, 50^{th}, and 97.5^{th} percentiles based on the AISW method with weights and scaling parameter shared across horizons with simple imputation (`AISW 2`) at the national level. Similarly to Figure 5, results for the period preceding the actual evaluation period are greyed out. As a remark, due to the introduced scaling parameter ϕ^{α}, the weights are not required to sum up to 1. The horizontal dashed line represents `weight` $= 1$.

4.4.5 Top-n model selection

Lastly, we consider the ensembles based on selection rather than weighting of members. As the user in practice needs to specify the number n of maintained models in advance, we assess the performance for all values $n = 1, \ldots, 8$ (with $n = 1$ corresponding to the selection of the top model only, and $n = 8$ corresponding to the unweighted ensemble).

In Figure 6, we show the results for $n = 4$, i.e., at each time point the better half of the models (over the training period) is included in the ensemble, with selection performed separately per horizon (`Select-4-Mean2` and `Select-4-Median2`). A graphical illustration of the respective nowcasts has been included in the bottom row of Figure 8. Despite some visually discernible differences to the unweighted ensembles (left panel), the average WIS values of `Select-4-Mean2` remain very close to those of the unweighted ensemble. Interval coverage rates are again somewhat improved. Figure 11 shows the overall WIS for the different values of $n = 1, \ldots, 8$ and the mean (left panel) and median (right panel) as the combination function. Red dots represent the results when the set of n models is updated every day, as would be done in a real-time application. For context, we show the results for all possible combinations of n models, keeping the selection of models constant over time, horizons and quantiles. Several conclusions can be drawn from the plot. Firstly, performance overall improves the more models are included into the ensemble, and only few model combinations at $n = 3$ through 7 achieve slight improvements over the full ensemble with $n = 8$. On the other hand, selection in real time (red dots) is always quite close to the optimum that could be achieved with a time-constant model selection, and comes close to the full unweighted ensemble from $n = 3$ onwards.

In the Supplementary Material, we present the corresponding results for the settings where the models are chosen jointly for all horizons (`Select-n-Mean1` and `Select-n-Median1`). Performance is overall somewhat weaker than when selection is done separately per horizon.

While again there is no clear improvement over the unweighted ensemble, our results indicate

that the effort necessary to maintain an ensemble model with numerous members may be reduced by restricting it to a few well-chosen members after an initial performance assessment.

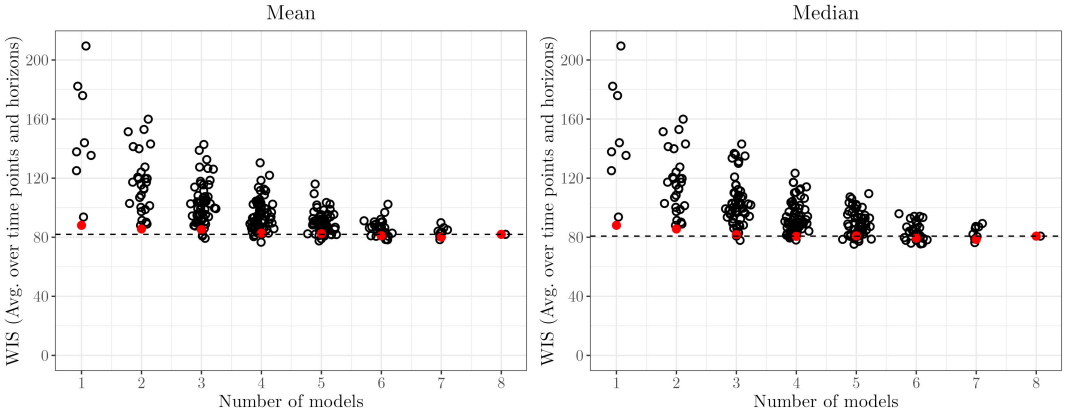

Figure 11: WIS (averaged over time points and horizons) for $n = 1, \ldots, 8$ in the `Select-n-Mean2` (left) and `Select-n-Median2` (right) models. Red circles show results for model selection updated each day, as would be done ine a real-time setting. For context, black circles show average values for all possible combinations of models when keeping the selection fixed over time. The horizontal dashed line represents the average WIS achieved by the full ensemble with all eight member models.

5 Discussion

In this paper, we proposed and analyzed different post-processing and ensemble techniques for the nowcasting of infectious diseases. In an application to COVID-19 hospitalization numbers from Germany we found that post-processing of individual models yielded performance gains across almost all considered models and technical specifications. This held both in terms of average WIS values and nowcast interval coverage. Somewhat surprisingly, post-processing of unweighted ensemble nowcasts did not yield improved performance, nor did post-processing of members prior to ensembling. More generally, it proved very challenging to improve upon unweighted mean and median ensembles. A straightforward direct inverse-score weighting approach led to very minor improvements, while a more sophisticated approach with weights optimized based on recent nowcast and observation pairs actually led to a decline in performance. Data-driven restriction of the ensemble to models with good recent performance did not yield improved performance either. However, the results indicate that the size of the ensemble, and thus the effort needed to maintain it, can be reduced without major losses in performance.

In the present paper we attempted to cover a spectrum of methods of moderate complexity which could be employed in practice. Many other extensions and alternative approaches could be explored. However, our general takeaway is that added complexity did not translate to improved performance. Some more flexible approaches we explored, like quantile regression with unconstrained weights for each model, were not tractable in our setting. Successful weighting approaches may thus need to be more parsimonious rather than more complex than what we assessed.

The fact that improved calibration (interval coverage) of post-processed and weighted ensembles did not yield improved performance in terms of average WIS may also reflect that the WIS and the CRPS, which it approximates, are relatively insensitive to overconfidence of predictions (see discussion in Bracher et al. 2021a). In terms of e.g., the logarithmic score, the more dispersed weighted ensembles might well yield improvements. However, the logarithmic score cannot be evaluated for forecasts stored as quantiles, and its practical application is often hampered by its extreme sensitivity to occasional mis-

guided forecasts. We therefore focused on the WIS, which has become a standard metric in collaborative disease forecasting.

More or less sophisticated weighting schemes being unable to outperform simple unweighted ensembles is a common finding in the literature, and Stock and Watson (2004) have coined the term "forecast combination puzzle" for this phenomenon. Various theoretical and empirical arguments have been brought forward to explain it (e.g., Claeskens et al. 2016; Smith and Wallis 2009). The essence of these is that estimated weights are often poorly identified and quite variable. This has a negative effect on performance, which may exceed the cost of the bias inherent in uniform weighting. Estimation of weights is thus less promising the closer the "true" weights are to uniformity.

For the post-processing task, on the other hand, there were sufficiently pronounced patterns in historical nowcast errors to meaningfully improve performance. For most models, this concerned mainly an overconfidence issue, i.e., too narrow prediction intervals. Future users of nowcasting models should keep this in mind, and may want to include simple post-processing steps into their analysis pipelines.

An issue not addressed in the present paper is that of missing nowcast submissions. In the German hospitalizations nowcasting project, considerable effort was spent on collating a complete set of nowcasts, but in most practical settings this cannot be ensured. The considered post-processing and combination methods were chosen such that they can relatively easily be extended to account for missing submissions (see Ray et al. 2023). More flexible techniques like unconstrained quantile regression, on the other hand, typically struggle to accommodate this aspect.

Declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability statement

The nowcast data for all individual models are available at `https://github.com/KITmetricslab/hospitalization-nowcast-hub`. The code used to reproduce the results presented throughout this paper is available at `https://github.com/avramaral/ensemble_learning`.

Acknowledgements

The authors would like to thank all contributors to the German COVID-19 Hospitalization Nowcast Hub. André Ribeiro Amaral acknowledges support from Karlsruhe Institute of Technology via the Aspirant Postdoc Grant. Daniel Wolffram and Johannes Bracher were supported by the German Federal Ministry of Education and Research (BMBF) via the project RESPINOW. Daniel Wolffram was moreover supported by the Helmholtz Association under the joint research school HIDSS4Health – Helmholtz Information and Data Science School for Health. Johannes Bracher was moreover supported by the German Research Foundation (DFG), project 512483310.

References

Abbott, S., Lison, A. and Funk, S. (2021). Epinowcast: Flexible hierarchical nowcasting. `https://zenodo.org/record/7924463`.

an der Heiden, M. and Hamouda, O. (2020). Schätzung der aktuellen Entwicklung der SARS-CoV-2 Epidemie in Deutschland – Nowcasting. Epidemiologisches Bulletin *2020*, 10–15.

Bastos, L. S., Economou, T., Gomes, M. F. C., Villela, D. A. M., Coelho, F. C., Cruz, O. G., Stoner, O., Bailey, T. and Codeço, C. T. (2019). A modelling approach for correcting reporting delays in disease surveillance data. Statistics in Medicine *38*, 4363–4377.

Beesley, L. J., Osthus, D. and Del Valle, S. Y. (2022). Addressing delayed case reporting in infectious disease forecast modeling. PLOS Computational Biology *18*, e1010115.

Bracher, J., Ray, E. L., Gneiting, T. and Reich, N. G. (2021a). Evaluating epidemic forecasts in an interval format. PLOS Computational Biology *17*, e1008618.

Bracher, J., Wolffram, D., Deuschel, J., Görgen, K., Ketterer, J. L., Ullrich, A., Abbott, S., Barbarossa, M. V., Bertsimas, D., Bhatia, S. et al. (2021b). A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. Nature Communications *12*, 5173.

Claeskens, G., Magnus, J. R., Vasnev, A. L. and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting *32*, 754–762.

Codeco, C., Coelho, F., Cruz, O., Oliveira, S., Castro, T. and Bastos, L. (2018). Infodengue: A nowcasting system for the surveillance of arboviruses in Brazil. Revue d'Épidémiologie et de Santé Publique *66*, S386.

Cox, D. R. and Medley, G. F. (1989). A process of events with notification delay and the forecasting of AIDS. Philosophical Transactions of the Royal Society of London. B, Biological Sciences *325*, 135–145.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J. 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*, e2113561119.

DelSole, T., Nattala, J. and Tippett, M. K. (2014). Skill improvement from increased ensemble size and model diversity. Geophysical Research Letters *41*, 7331–7342.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. Environmetrics *23*, 579–594.

German Federal Government (18 November 2021 [Cited 19 July 2023]). Videoschaltkonferenz der Bundeskanzlerin mit den Regierungschefinnen und Regierungschefs der Länder am 18. November 2021. `https://www.bundesregierung.de/resource/blob/974430/1982598/defbdff47daf5f177586a5d34e8677e8/2021-11-18-mpk-data.pdf`.

German Federal Ministry of Health (7 October 2021 [Cited 19 July 2023]). FAQ zur Hospitalisierungsinzidenz. `https://www.bundesgesundheitsministerium.de/coronavirus/hospitalisierungsinzidenz.html`.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association *102*, 359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review *133*, 1098–1118.

Greene, S. K., McGough, S. F., Culp, G. M., Graf, L. E., Lipsitch, M., Menzies, N. A. and Kahn, R. (2021). Nowcasting for real-time COVID-19 tracking in New York City: an evaluation using reportable disease data from early in the pandemic. JMIR Public Health and Surveillance *7*, e25538.

Günther, F., Bender, A., Katz, K., Küchenhoff, H. and Höhle, M. (2021). Nowcasting the COVID-19 pandemic in Bavaria. Biometrical Journal *63*, 490–502.

Heyder, S. and Hotz, T. (2021). The ILM-prop model: code and explanation. `https://github.com/Stochastik-TU-Ilmenau/ILM-prop`.

Höhle, M. and an der Heiden, M. (2014). Bayesian nowcasting during the STECO104:H4 outbreak in Germany, 2011. Biometrics *70*, 993–1002.

Lison, A., Abbott, S., Huisman, J. and Stadler, T. (2024). Generative Bayesian modeling to nowcast

the effective reproduction number from line list data with missing symptom onset dates. PLOS Computational Biology *20*, 1–32.

Norddeutscher Rundfunk (20 November 2021 [Cited 19 July 2023]). Nach MPK-Beschluss: Verwirrung um Hospitalisierungsinzidenz. `https://www.ndr.de/nachrichten/info/Nach-MPK-Beschluss-Verwirrung-um-Hospitalisierungsinzidenz,hospitalisierungsinzidenz100.html`.

Ray, E. L., Brooks, L. C.and Bien, J., Biggerstaff, M., Bosse, N. I., Bracher, J. et al. (2023). Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. International Journal of Forecasting *39*, 1366–1383.

Reich, N. G., McGowan, C. J., Yamana, T. K., Tushar, A., Ray, E. L., Osthus, D. et al. (2019). Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the US. PLOS Computational Biology *15*, e1007486.

Reis, J., Yamana, T., Kandula, S. and Shaman, J. (2019). Superensemble forecast of respiratory syncytial virus outbreaks at national, regional, and state levels in the United States. Epidemics *26*, 1–8.

Schneble, M., De Nicola, G., Kauermann, G. and Berger, U. (2021). Nowcasting fatal COVID-19 infections on a regional level in Germany. Biometrical Journal *63*, 471–489.

Schulz, B., El Ayari, M., Lerch, S. and Baran, S. (2021). Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting. Solar Energy *220*, 1016–1031.

Seaman, S. R., Samartsidis, P., Kall, M. and De Angelis, D. (2022). Nowcasting COVID-19 deaths in England by age and region. Journal of the Royal Statistical Society Series C: Applied Statistics *71*, 1266–1281.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics *71*, 331–355.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. Journal of Forecasting *23*, 405–430.

Thorarinsdottir, L. T., Gneiting, T. and Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. SIAM/ASA Journal on Uncertainty Quantification *1*, 522–534.

van de Kassteele, J., Eilers, P. H. C. and Wallinga, J. (2019). Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing. Epidemiology (Cambridge, Mass.) *30*, 737.

Wolffram, D., Abbott, S., an der Heiden, M., Funk, S., Günther, F., Hailer, D. et al. (2023). Collaborative nowcasting of COVID-19 hospitalization incidences in Germany. PLOS Computational Biology *19*, 1–25.

Yamana, T. K., Kandula, S. and Shaman, J. (2016). Superensemble forecasts of dengue outbreaks. Journal of The Royal Society Interface *13*, 20160410.

Zamo, M., Bel, L. and Mestre, O. (2021). Sequential aggregation of probabilistic forecasts–application to wind speed ensemble forecasts. Journal of the Royal Statistical Society Series C: Applied Statistics *70*, 202–225.

A Individual Models

A summary of characteristics of the eight individual models can be found in (Wolffram et al., 2023, Table 1). We here only provide a high-level summary for each model, along with the reference to the respective original paper or documentation.

1. `Epiforecasts` (Abbott et al., 2021): A Bayesian approach combining flexible modelling of delay distributions (including weekday effects) and a random walk prior on the latent time series of complete hospitalization counts.

2. `ILM` (Heyder and Hotz, 2021): Predictions of yet unreported hospitalizations are based on case counts, for which suitable multiplication factors are derived per age group. Uncertainty intervals are based on past nowcast errors. Note that we here do not use the model as run in real time, but a retrospectively re-run version with an assumed maximum delay of 42 days. The real-time model featured a maximum delay of 80 days, which was not suitable for the purposes of the present study.
3. `KIT` (Wolffram et al., 2023, Supplementary Section E): A simple multiplication factor model ignoring weekday effects. Uncertainty intervals are based on past nowcast errors.
4. `LMU` (Schneble et al., 2021): A frequentist nowcasting approach combining a generalized additive model for the reporting triangle and a sequential multinomial model for the delay distribution.
5. `RIVM` (van de Kassteele et al., 2019): The reporting triangle is modelled using a bivariate spline surface, accounting for weekday effects. This is used to extrapolate to the unobserved parts of the triangle.
6. `RKI` (an der Heiden and Hamouda, 2020): Logistic regression is used to model conditional reporting probabilities, taking into account various covariates (weekdays, states, age groups, etc.).
7. `SU` (Günther et al., 2021): Conceptually close to Epiforecasts, combines a random walk prior with a discrete-time hazard model for the reporting delay.
8. `SZ`: Nowcasts are based on empirical ratios of preliminary and completed data, for which various quantiles are computed.

B Decomposition of the weighted interval score

Rather than via the linear quantile score, the weighted interval score can also be defined as a weighted sum of interval scores (thus its name). The interval score for a prediction interval $[l_\beta, u_\beta]$ at nominal coverage level $(1-\beta)$ is given by (Gneiting and Raftery, 2007)

$$\text{IS}_\beta(l_\beta, u_\beta, x) = (u_\beta - l_\beta) + \frac{2}{\beta} \times (l_\beta - x) \times \mathbb{1}(x < l_\beta) + \frac{2}{\beta} \times (x - u_\beta) \times \mathbb{1}(x > u_\beta).$$

It obviously consists of three components, namely

1. a component for forecast dispersion, given by the interval width $u_\beta - l_\beta$.
2. a component for overprediction, i.e. a penalty in the case $x > u$.
3. a component for underprediction, i.e., a penalty in the case $x < l$.

Now assume that A is uneven and that the quantile levels are chosen such that $q^{\alpha_1}, \ldots, q^{\alpha_A}$ correspond to the predictive median $m = q^{\alpha_{(A+1)/2}}$ and the ends $l_{\beta_1} = q^{\alpha_1}, \ldots, l_{\beta_{(A-1)/2}} = q^{\alpha_{(A-1)/2}}, u_{\beta_1} = q^{\alpha_{(A+1)/2+1}}, \ldots, u_{\beta_{(A-1)/2}} = q^{\alpha_A}$ of $(A-1)/2$ nested central prediction intervals. These intervals have nominal coverage levels $(1-\beta_j) = \alpha_{A+1-j} - \alpha_j, j = 1, \ldots A$. The weighted interval score can then also be written as (Bracher et al., 2021a)

$$\text{WIS}(l_{\beta_1}, \ldots, l_{\beta_{(A+1)/2}}, m, u_{\beta_{(A+1)/2}}, \ldots, u_{\beta_1}, x) = \frac{1}{A} \times \left(|x - m| + \sum_{j=1}^{(A-1)/2} \beta_j \times \text{WIS}(l_{\beta_j}, u_{\beta_j}, x) \right). \quad (10)$$

It thus inherits the decomposition of the interval score. In practice, the components can also be computed in a more straightforward way without recurring to the somewhat involved formulation (10). This can be done as follows:

1. The spread component equals the WIS with the observation x replaced by the predictive median m.
2. The overprediction component is zero if $x < m$. If $x > m$, it is given by the difference of the WIS and the spread component.

3. Accordingly, the underprediction component is zero if $x > m$. If $x < m$ it is given by the difference of the WIS and the spread component.

C The approximate integrated quadratic distance

The simple imputation approach, where preliminary observations are replaced by up-to-date predictive medians in forecast evaluation, neglects that there is uncertainty attached to these imputed values. It would be desirable to account for this and take into account the full current predictive distribution (or, in our setting, all available quantiles). To this end we use an approximation of the integrated quadratic distance (Thorarinsdottir et al., 2013), which we motivate below.

As mentioned in the main manuscript, the WIS is a quantile-based approximation of the continuous ranked probability score (CRPS). For a predictive cumulative distribution function F and observed value x this score is defined as

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} [F(z) - \mathbb{1}(z \geq x)]^2 \mathrm{d}z.$$

For the CRPS, a setting similar to ours is mentioned by Friederichs and Thorarinsdottir (2012). They address the evaluation of forecasts for quantities observed with an observation error. Denoting by x the observed value including the observation error, they propose to use the score (see also Brehmer and Gneiting 2019)

$$S(F, x) = \int_{-\infty}^{\infty} [F(z) - \Phi(z - x)]^2 \mathrm{d}z,$$

where Φ is the cumulative density function assumed for the observation error. This can be read as a comparison of two probability density functions: the one of the forecast, and an assumed distribution of the true value given the imperfect observation x. The resulting distance between the two cumulative distribution functions is called the *integrated quadratic distance* (IQD) (Thorarinsdottir et al., 2013).

This fits our setting well, as we likewise wish to compare one predictive density to an "uncertain" observation described by another probability distribution. Denoting the CDF of the nowcast distribution to evaluate by F and the CDF of the most recent nowcast used as the "uncertain truth" by F^*, we could therefore use the integrated squared distance

$$\mathrm{IQD}(F, F^*) = \int_{-\infty}^{\infty} [F(z) - F^*(z)]^2 \mathrm{d}z$$

for evaluation. This, of course, is not feasible in practice as we only know a set of predictive quantiles for both F and F^*. We denote these quantiles by $q_F^{\alpha_1}, \ldots, q_F^{\alpha_A}$ and $q_{F^*}^{\alpha_1}, \ldots, q_{F^*}^{\alpha_A}$, respectively. It can be demonstrated that an approximation of the IQD paralleling the approximation of the CRPS by the WIS is (Resin et al 2024)

$$\mathrm{IQD}(F, F^*) \approx \sum_{i=1}^{A} \sum_{j=1}^{A} \frac{\alpha_{i+1} + \alpha_{i-1}}{2} \times \frac{\alpha_{j+1} + \alpha_{j-1}}{2} \times \chi(\alpha_i, \alpha_j, q_F^{\alpha_i}, q_{F^*}^{\alpha_j}) \times \left| q_F^{\alpha_i} - q_{F^*}^{\alpha_j} \right|,$$

where

$$\chi(\alpha_i, \alpha_j, q_F^{\alpha_i}, q_{F^*}^{\alpha_j}) = \begin{cases} 1 & \text{if } (\alpha_i > \alpha_j \text{ and } q_F^{\alpha_i} < q_{F^*}^{\alpha_j}) \text{ or } (\alpha_i < \alpha_j \text{ and } q_F^{\alpha_i} > q_{F^*}^{\alpha_j}) \\ \frac{1}{2} & \text{if } \alpha_i = \alpha_j \\ 0 & \text{else.} \end{cases}$$

We thus use this expression to evaluate nowcast quantiles against preliminary observations ac-

counting for the remaining uncertainty. When evaluating the performance separately per quantile level α_i of F, we can just use the summands for that α_i, just like we are only using the respective linear quantile score rather than the full WIS if the observed value x is known with certainty.

Additional references

Brehmer, J. and Gneiting, T. (2019). Properization: constructing proper scoring rules via Bayes acts. Annals of the Institute of Statistical Mathematics 72, 659–673.

Resin, J., Bracher, J., Dimitriadis, T. and Wolffram, D. (2024). Dispersion-shift decompositions of Wasserstein and Cram ér Distances. In preparation.