GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA
SEIT 1737

Analyzing the Influence of Model and Ensemble Structure on Performance of Real-Time COVID-19 Forecasts

Master thesis for the Master of Science course Applied Statistics
at the University of Göttingen

Author:
Friederike Sarah Becker,
Student ID: 21914687,
born in Herten, Germany

Supervisors
Prof. Dr. Thomas Kneib
Nikos Bosse

Submitted on September 22, 2022
to the Faculty of of Business and Economics at Göttingen University

Contents

List of Figures

List of Tables

Acronyms

AE absolute error

CDC Centers for Disease Control and Prevention

CRPS continuous ranked probability score

ECDC European Center for Disease Prevention and Control

EW Epidemiological Week

GLM generalized linear model

JHU Johns Hopkins University

LSHTM London School of Hygiene and Tropical Medicine

QRA quantile regression average

WIS weighted interval score

Acknowledgments

I would like to thank my supervisors, Nikos Bosse and Thomas Kneib, for their continuous support, guidance and advice. Additionally, to Professor Kneib: I am thankful first and foremost for his teaching, but also for the kindness and flexibility he has shown towards me and other students throughout the master in Applied Statistics. To Nikos, thank you for giving me the opportunity to work on this project, and for your support, both personally and academically.

Thank you to Sebastian Funk for welcoming me into the Epiforecasts working group at the London School of Hygiene and Tropical Medicine, as well as for giving me valuable advice and feedback on my work. Likewise, I am thankful to Kath, Hannah, Nikos, Sophie, Hugo, Akira, James and Sam for welcoming me and providing both direct feedback and inspiration to this work as well as countless valuable learnings on topics related to academia and research.

Thank you to Lukas, Jo and Marina for reading parts of this work and providing feedback.

Thank you to my friends and family for their support, both during this project and the many years before.

Eigenständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die von mir angegebenen Quellen und Hilfsmittel verwendet habe.
Wörtlich oder sinngemäß aus anderen Werken entnommene Stellen habe ich unter Angabe der Quellen kenntlich gemacht.
Die Richtlinien zur Sicherung der guten wissenschaftlichen Praxis an der Universität Göttingen wurden von mir beachtet.
Mir ist bewusst, dass bei Verstoß gegen diese Grundsätze die Prüfung mit nicht bestanden bewertet wird.

(Quelle Eigenständigkeitserklärung:

https://www.uni-goettingen.de/de/document/download/cdc9faa5ff3a71599b0b2b9e7a3d1eb4.pdf/FN2_Selbstaendigkeitserklaerung.pdf)

Analysis and Code

The analyses underlying this work were performed in R (R Core Team, 2022).
Code and data are publicly available under https://github.com/fredbec/masterthesis. The dataset was originally obtained from https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe.
Scoring of forecasts was generally conducted with the `scoringutils` package (Bosse et al., 2022a).
Use of implemented methodology in other packages (with the exception of functions from the tidyverse) is generally noted separately.

1 Introduction

In recent years evidence in epidemiology - as well other fields - has accumulated that in order to obtain accurate and well calibrated forecasts of targets of interest, such as case numbers of a disease in a given week, it is often advisable to not rely solely on single model outputs, but to rather consider an aggregate of forecasts made by a group of models, widely referred to as ensemble forecasts (see for instance Yamana et al. (2016), Reich et al. (2019), McGowan et al. (2019)).
The recent COVID-19 epidemic has turned out to be no exception in this regard. In order to obtain an accurate picture of current epidemic dynamics for decision makers, researchers and the public alike, in March of 2021 the European Center for Disease Prevention and Control (ECDC) instigated the European COVID-19 Forecast Hub (Sherratt et al., 2022a), collating weekly real-time distributional forecasts for short-term incident COVID-19 cases and deaths from independent modeling teams. This followed after similar efforts in the U.S.[1] by the Centers for Disease Control and Prevention (CDC) (Cramer et al., 2021). It was generally found that aggregation of individual model outputs into a combined forecast showed more consistent performance than any or most individual models for forecasting both COVID-19 cases and deaths (e.g. Sherratt et al. (2022b) (European Hub), Cramer et al. (2022) (U.S. Hub)).
However, while empirical evidence for the advantages of employing an ensemble strategy is ubiquitous, a question that often remains in a given forecasting setting is whether certain types of model or ensemble compositions yield improved performance, as well as how ensembles interact with the models they are composed of. Within this work, the focus will lie on several dimensions of this issue.

For the goal of eliciting accurate short-term forecasts in epidemiology, it is an ongoing topic of research whether models that are to some extent explicitly epidemiological in their modeling strategy[2] yield higher predictive accuracy than models that are agnostic to the underlying transmission dynamics and thus tend to solely rely on past information of the target time series (Funk et al., 2020). Different diseases and epidemics might warrant varying approaches in modeling strategy and we thus investigate whether, for forecasts of COVID-19 in the European Hub, definitive or (more likely) situation-dependent rankings between different modeling strategies can be established, for example along the lines of different phases of the pandemic. Furthermore, there might be different requirements and goals for forecasts depending on who is using them and under which circumstances (Bosse et al., 2022b). Differences between modeling strategies thus might also show along the lines of the employed assessment method.

Furthermore, while it is a universal result that ensemble models generally exhibit more robust performance than single models, little is known about how individual models can affect an ensemble's performance. This can however be a relevant question: consider the situation where, given an already established ensemble, one is proposed a new model and is consequently confronted with the decision of whether to add it to the base ensemble. In principle, an extra model can have negative effects on an ensemble, presumably more for smaller ensembles that are not as robust to additions. The usual practice is often to simply include additional models, but we want to investigate how single models can impact the performance of an ensemble, particularly as this relates to the size of the ensemble, as well as the recent performance of individual models or their agreement with the predictions of the current ensemble.
Another central question in past and current ensembling efforts has been whether ensembles should discriminate between their potential member models based on merit and thus assign higher weights to models that have shown better historical performance than others. For COVID-19 forecasts, results on

[1] We will generally refer to the respective efforts as "U.S. Hub" and "European Hub", for the sake of brevity.
[2] By explicitly modeling disease transmission dynamics or being grounded in other epidemiological foundations

performance-based weighting have been somewhat mixed, with some positive results in the U.S. Hub (e.g. Taylor and Taylor (2021), Ray et al. (2022)), which could generally not be confirmed in European applications (e.g. Sherratt et al. (2022b), Bracher et al. (2021b)). Taylor and Taylor (2021) identify a possible reason for the shortcomings of weighted approaches: they require comparable records of historical accuracy and are thus challenging to implement in datasets where model availability fluctuates - this is presumably especially an issue for the European Hub, where participation gaps are common.
However, it might yet be possible to leverage recent differences in model performance for ensemble composition in the European Hub, for instance by first reducing the set of models considered for such a weighting scheme. Furthermore, we deem it interesting to investigate possible reasons for why a weighted approach might perform poorly and thus attempt to to establish some guidelines (for example pertaining to the specific target series or the available set of component models), for when it might be beneficial to select and/or weight models and when it might be better to rely on an aggregation method that treats all component models equally.
Furthermore, results that emerge from the analysis of modeling strategies could potentially also be leveraged for ensemble forecast composition - for instance, in their analysis of COVID-19 death forecasts in the U.S., Taylor and Taylor (2021) surmise that models which explicitly model transmission dynamics should in principle be better suited to issuing accurate forecasts in an epidemiological setting. They thereby construct ensembles that only consist of these, with partially positive results. We thus investigate whether we can similarly use the categorization of modeling strategies to improve ensemble performance.

In general, our analysis and procedure should, to some extent, be regarded more inquisitively and with the aim to establish heuristics rather than hard and fast rules. This is especially due to the fact that our analysis is entirely retrospective and emerging results might be very specific to the data at hand and thus not generalizable. Furthermore, we want to emphasize that our goal is not necessarily to find a new method that should directly replace the use of the current real-time ensemble in the European Hub. We rather ask ourselves: with mild hindsight knowledge, is it possible to beat the median ensemble, consistently over locations and different time periods? If yes, this would propose a new ensembling strategy that could be employed and assessed alongside the current ensemble method in real-time. If not, we have produced more evidence that supports the continued use of the median ensemble.

Concretely, the structure of this thesis is thus as follows: In section 2, we formally introduce the practice of forecasting, both generally as well as in the realm of epidemiology, specifically with respect to challenges in epidemiological forecasting and the different modeling strategies that are commonly used in the field. In Section 3, we introduce the central methodological concepts that underly this work, most importantly the methods and tools that are commonly used to assess and score forecast performance, as well as different aggregation techniques for combining component forecasts into an ensemble. In Section 4, we give a longer introduction of the European COVID-19 Forecast Hub dataset, including a discussion of some previous results for forecasting of COVID-19 and an explanation of the structure of the data subset we are using. Then, section 5 will follow with some exploratory analyses of model and ensemble behavior - concretely, we investigate performance differences of general modeling strategies and furthermore perform a systematic experiment that seeks to investigate how sensitive an unweighted mean or median ensemble is to model additions. Lastly, in section 6, we retrospectively apply ensembling methods that go beyond the direct weighting of all component models, with the goal of potentially identifying a method that could improve the performance of an ensemble model in the European Hub. Concretely, we attempt to improve performance by prior selection of better performers as well as weighting at the level of modeling strategies. Finally, section 7 closes with a summary and discussion of our results.

2 Forecasting and Ensembles

In this section, we introduce the concept of forecasting in general, as well as how it applies to the field of epidemiology - we briefly characterize potential difficulties of forecasting in this field and introduce general modeling strategies that are commonly used. Lastly, we introduce the concept and practice of ensemble forecasts.

2.1 (Epidemic) forecasting

We begin by defining the practice of forecasting, as we observe that it is sometimes conflated with related but distinct terms, such as the more general task of prediction. Moran et al. (2016) define forecasting as the “ability to predict what will happen in the future on the basis of analysis of past and current data”. Where prediction modeling is thus the practice of issuing categorical or quantitative statements for any type of unseen data, forecasting explicitly refers to the case where that unseen data lies in the future, and predictions are made based on historical data.

An important feature of a useful forecast is that it should be probabilistic in nature and thus take on the form of a predictive probability distribution (Gneiting et al., 2007). A probabilistic forecast is richer in information and gives the end user of the forecast a more realistic picture of what to expect - especially when planning for adverse outcomes, one can imagine that interest often lies less in a point prediction and more in anticipating and preparing for a range of plausible values that the outcome might attain.

A central feature of forecasts is that they are issued as explicit and non-contingent statements about a future quantity, thereby typically reducing the time horizon for useful forecasts (Reich et al., 2022). This is due to the fact that the relevant circumstances affecting said quantity need to be somewhat constant in order to make reliable forecasts. If the underlying system is more unstable, uncertainty about future outcomes is greater and the viable forecast horizon is consequently reduced. On the flip side, non-contingency means that forecasts can directly be evaluated against the truth data that realized, which facilitates model evaluation and development (Reich et al., 2022).

Epidemic forecasting then simply refers to the practice of issuing forecasts for relevant quantities that characterize the future course of an epidemic. In the context of COVID-19, forecast targets were usually counts of confirmed cases directly, or deaths or hospitalizations that resulted from infections. Examples of other targets in epidemic forecasting are the timing or height of a peak in observed infections - these targets are however more relevant for seasonal diseases such as influenza, see Reich et al. (2019) for an example.

There are a few characteristics that can make forecasting epidemics in comparison to other forecasting efforts particularly challenging. Firstly, Jajosky and Groseclose (2004) state that data sources are often not available in real-time, and are sometimes made available with substantial lag after initial recording. In the context of COVID-19, this was admittedly less of an issue, as data were often assembled and made available - to researchers and the public alike - in real-time, most notably by Johns Hopkins University (JHU) (Dong et al., 2020). Nevertheless, these data often saw subsequent revisions and could therefore to a certain degree be unreliable at time of release (Sherratt et al., 2022a).

Another factor that makes forecasting epidemics challenging is that the dynamics of disease outbreaks are affected by a host of factors that can be difficult or impossible to fully represent in forecast models and that are additionally subject to constant change, such as human behavior and/or the characteristics of the relevant pathogen (Moran et al., 2016). Specifically in the context of COVID-19, Cramer et al. (2022) mention that epidemic forecasts can also play a role in impacting human behavior (by affecting policy and/or individual behavior directly), thereby creating a sort of “feedback loop” that forecasters would additionally need to account for. These challenges also relate back to the previous point about the limited time horizon of useful forecasts - uncertainties about changes in the epidemic process, such as the potential emergence of new variants, or in human behavior, for instance induced by changes in policy, limit forecasts’ reliability and horizons in the context of COVID-19 were generally limited to a few weeks in most forecasting efforts (Reich et al., 2022).

If interest instead lies in longer term projections, one can pre-define settings for the relevant circumstances and thereby elicit plausible future trajectories, a practice known as scenario modeling (Reich et al., 2022). Scenario modeling can be used to inform decision makers that seek to evaluate the plausible effects of potential strategies, e.g. specific disease control measures, under a variety of circumstances (Reich et al., 2022). Contrary to forecasts, they can not be reliably tested against truth data, due to the fact that it is highly unlikely that the circumstances realize in the exact way they were defined. Scenario modeling is however not the focus of this work, so we do not expand this further.

In the following, we introduce some modeling strategies that are commonly used for short-term forecasting of infectious diseases and constitute the majority of models in the European Hub data.

Mechanistic - alternatively named compartmental - models are among the most widely used approaches in epidemiology to model infectious diseases. The term “compartmental” derives from the fact that these models divide the population into compartments according to their infection status with respect to the infectious disease at time t. In their most basic structure, these compartments are S, I and R (Brauer and Castillo-Chavez, 2012): $S(t)$ denotes the number of individuals that are susceptible to the disease at time t, that is, they can potentially be infected with the disease. $I(t)$ denotes the number of infected individuals, which are assumed to be able to spread the disease when in contact with individuals from compartment S. Finally, $R(t)$ denote the number of individuals that, after being infected, have been removed from the process at time t. This compartment thus contains individuals that are either isolated, recovered/immunized without possibility of reinfection, or dead as a consequence of the disease. Additional characteristics of the epidemiological process of a particular infectious disease, for instance the availability of a vaccine, can be modeled via extra compartments. Commonly added extra compartments are E for exposed but not yet infectious individuals, giving an SEIR model, or an additional S compartment to model non-permanent immunity from the disease, resulting in an SIRS model (Brauer and Castillo-Chavez, 2012). Given some regularity assumptions and a set of parameters governing the transmission process, flow of individuals between the compartments is then modeled.

Furthermore, there exist models that are built on epidemiological information but do not have an explicit compartmental framework - these models can be grouped together in a *semi-mechanistic* category. This category most notably includes models that estimate the time-varying reproduction number R_t (defined as the average number of secondary infections caused by an infected individual at time t) and subsequently translate this to number of infections via e.g. the renewal equation (Fraser, 2007), or directly estimate the epidemic’s growth rate and map this to new infections. Deaths can either be modeled as a separate process or be mapped from estimated or observed cases.

Another strategy for forecasting in epidemiology are *statistical* models, which refers to models that in their modeling approach are agnostic to the underlying transmission dynamics. The term thus derives from the fact that models from this category rely *solely* on statistical methods (usually based on past data of the time series) to make predictions - put succinctly, Holmdahl and Buckee (2020) refer to this strategy as “crunch[ing]” epidemiological data from the past [...] and project[ing] cases into the future”. These models are mostly time series - for example ARIMA - models.

While models of these types constitute the bulk of models used in applications (both in this and in other works we consulted), the list is not entirely exhaustive. There also exist agent-based models, which model epidemiological processes by simulating behavior of each individual in a population - Zelner et al. (2021) state that these can provide useful insight but can be very difficult to fit to data. Yet another strategy results from directly aggregating forecasts based on human judgment (either from a number of experts within the field or members of the general public).

These approaches are all distinct in the way that they approach the task of issuing short-term forecasts, but there is no consensus in the epidemic modeling community on which of them performs best (Moran et al., 2016). Before eventually turning to our analysis of model strategy performance in our dataset in section 5.1, we first want to highlight some previous findings in the context of other diseases, as well as discuss some reasonings that argue for or against the different modeling strategies on a purely theoretical basis.

For the case of influenza, Reich et al. (2019) found that statistical models exhibited similar or better performance than mechanistic models in short-term forecasts, especially at longer horizons, while McGowan et al. (2019) similarly found that statistical models outperformed mechanistic models, thereby suggesting that explicitly modeling the epidemiological process does not necessarily provide an advantage in short-term predictions.

However, with regard to these findings, Bracher et al. (2021b) argue that within the setting of emerging diseases such as COVID-19 (rather than seasonal diseases such as influenza), mechanistic models might have an advantage due to the limited amount of historical data: while statistical models can predict the seasonal patterns of influenza given data on previous outbreaks, emerging diseases generally do not neatly follow seasonal patterns. In particular, it may be harder for statistical models to accurately predict the effect of new interventions, as they have no past data to base the estimation on. Contrary to this, mechanistic models have key parameters with an interpretable role in the context of the transmission system, which researchers can tune to account for e.g. interventions.

On a more conceptual level, as mechanistic models explicitly attempt to model underlying transmission dynamics, they can generate insights of the process that purely statistical models cannot (James et al., 2021). They are thus also suited to longer term planning via scenario modeling, and thus for instance gauge the expected effect of a proposed policy measure, while the application of statistical models is purely limited to shorter term predictions (Reich et al., 2022). On the flip side, this can also mean that accuracy for models that explicitly account for epidemiological information is also somewhat limited by knowledge about the virus (Holmdahl and Buckee, 2020).

In regards to forecasts based on human judgment, Bosse et al. (2021) saw promising results from such an approach within the context of predicting COVID-19 case numbers: centrally, forecasts based on human judgment are thought to be better able to anticipate changes in trends, as they are, for instance, able to incorporate knowledge about changes in policy in a way more rigid modeling strategies are not. This can be especially valuable due to the fact that, in the context of COVID-19, other models have often been found to be less able at predicting changes in trends, especially for forecasting case numbers (Ray et al., 2021).

Within the context of this work, we categorized models based on the metadata files that modeling teams submitted along with their forecasts - these files gave details about the model and fitting process among other things, such as the truth data source they used to train their model on. Furthermore, for many modeling teams, additional resources could be found, such as websites or their own publications about the model, which we also consulted.[3]

We also want to mention that distinction between model types may not always be entirely clear. For instance, as noted by Reich et al. (2019), mechanistic models often do have statistical components to account for systematic discrepancies between noisy surveillance data and the actual transmission numbers these models seek to predict - they decide to categorize any model as mechanistic which made a mention of having an explicit compartmental framework, which we follow. Furthermore, we stay closely aligned to Bracher et al. (2021b), who largely used the same categories when sorting their models (although they used different names for their categories). Within our dataset, the availability of agent-based models and expert judgment models was more limited than that of other modeling strategies. In particular, the expert judgment based approach was only available for roughly the first half of the study period, while agent-based models were only available for one location within our subset of the data. We thus mainly focus on the models in the categories mechanistic, semi-mechanistic and statistical.

2.2 Ensemble Forecasts

Rather than relying on a single model's output for forecasting, the practice in many fields nowadays is to instead employ aggregation strategies that unify a number of predictions into a single forecast, thereby aiming to reduce variability in individual model performance and ideally also take advantage of respective models' strengths, be it superior modeling strategy or a better information set. This practice is known as building ensemble forecasts.

Basic intuition for why combining a set of forecasts rather than relying on a single forecasting model can be beneficial derives from the fundamental statistical fact that averaging over a set of independent unbiased estimates for a quantity still gives an unbiased estimate of that quantity, but with reduced variance. That is, through combining forecasts that we believe to on average yield accurate predictions for the quantity of interest, we can to a certain extent cancel out the forecast errors they individually make at a given time point in practice - this of course assumes that the forecast errors they make are not correlated.

While these conditions of unbiasedness and uncorrelatedness are of course unlikely to completely hold in practice, Timmermann (2006) argues that even in these cases, combining forecasts still provides benefits over relying on a single forecast as a general practice. For instance, forecasts might be individually biased at certain time points, possible reasons for which might be model misspecification or data errors. Such a bias might however be hard to detect in real-time, making ensemble forecasts a robust mitigation strategy, as other models presumably did not suffer from the same issues and can counterbalance these predictions. In case of correlated forecast errors, for instance following an unforeseen intervention, Timmermann (2006) similarly argues that combination methods can guard against deteriorations of component forecast performance. In fact, as single component forecasts presumably show the largest variation in these cases, the benefit of combination can also presumed to be the greatest in the presence of structural breaks.

These theoretical arguments are backed up by empirical results - in practice, ensemble forecasts are widespread and usually show superior performance to forecasts made by individual models. This result holds across fields, e.g. weather forecasting (Krishnamurti et al., 2000) and also epidemic forecasting: in the context of the latter, ensemble forecasts showed superior performance for influenza (e.g. Reich et al.

[3] The spreadsheet/csv file detailing the categorizations can be found in the Github repository of this thesis, in the 'scraper' directory.

(2022)), dengue (e.g. Johansson et al. (2019)) and also recently for COVID-19 (for instance Cramer et al. (2022), Sherratt et al. (2022a)). In some applications, it was found that a small number of top-performing models outperformed the ensemble, for instance by McGowan et al. (2019) (influenza) and Bracher et al. (2021b) (COVID-19) - nevertheless, even in these cases, ensemble performance was generally still robust and an improvement upon most individual models. Moreover, improvements could even be achieved when models did not originate from several teams, but rather from a single team incorporating multiple in-house models, where diversity in knowledge and strategies is presumably more limited (e.g. Reich et al. (2019)).

However, a question that often remains is which concrete method to use for combining forecasts into an ensemble. In particular, it is often nor clear whether to consider all models equally or instead give models with a proven track record of good performance higher weight. In principle, it would be desirable to assign higher weights to models that show increased skill, as this practice could for instance also leverage individual model strengths, such as different information sets. In practice, ensemble forecasts with estimated weights often perform poorly relative to unweighted aggregation methods, a fact that came to be known as the "forecast combination puzzle" - see Claeskens et al. (2016) for a theoretical argument for why this might be the case.

Within the context of COVID-19, results on this question have been somewhat varied. In an application in the U.S., Ray et al. (2022) saw improved performance for forecasting deaths from weighting approaches. In a related European application, these results could not be confirmed (Sherratt et al., 2022a). We however only make short mention of these results here, as we will extensively discuss in our own application in section 6.1. Within the next section, we will also go into detail on different techniques to combine forecasts into an ensemble, in particular methods that can be used to weight component forecasts based on their past performance.

3 Methodology

In this section, we introduce the central methodological concepts that underly this work, mainly those of assessing forecast performance. We thus intend to give an overview of the toolkit that is available to assess whether forecast models give accurate predictions and which we use within the context of this work. We first define the general format of predictive distributions that we will assess, then introduce the forecasting paradigm and subsequently the metrics and proper scoring rules that rely on this notion. Additionally, we explain characteristics of these metrics and give short examples of how to interpret them. Furthermore, we formally introduce the methods that can be used to build ensemble predictions and close with a short note on analysis methods and statistical testing for forecasts.

3.1 Format of forecast predictions

The forecasts that are analyzed in this thesis are issued as predictive distributions, in the form of a set of 23 discrete and unequally-spaced quantiles. Formally, we write that the forecast distribution for model m at time [4] t $F_{m,t}$ is comprised of a set of $K = 23$ quantiles for horizons h between one to four weeks ahead: $q_{m,t,h,\tau}$, where $\tau \in \{0.01, 0.025, 0.05, 0.1, ...0.9, 0.95, 0.975, 0.99\}$. Note that quantile levels are not equally spaced, giving the predictive distributions a higher resolution around their respective tails. These quantile levels directly induce 11 central prediction intervals with nominal coverage level $(1 - \alpha)\%$, with $\alpha \in \{0.02, 0.05, 0.1, 0.2, ...0.9\}$.

As this is the common format of issuing predictive distributions within the literature of forecasting COVID-19, methods of assessing and scoring model performance have been readily developed from more common formats into the setting of quantile-based predictive distributions. Before we turn to concretely discussing these methods, we briefly introduce the concept of sharpness subject to calibration, that is, the forecasting paradigm that is at the basis of the way forecasts are scored and assessed in most applications.

3.2 Assessing forecast performance

3.2.1 The forecasting paradigm

To better understand what an optimal forecast looks like and, in the same vein, how to best assess forecast performance, we follow the established and ubiquitous paradigm characterized by Gneiting et al. (2007): an optimal forecasts should maximize *sharpness subject to calibration.*

Consider that we aim to make a probabilistic forecast F_t for a quantity y_t at time t, which follows the distribution G_t. The ideal forecaster would thus issue

$$F_t = G_t$$

as their probabilistic forecast. In practice, y_t will eventually be observed, while the data-generating distribution G_t remains an unknown and hypothetical theoretical concept. The skill of the forecaster thus needs to be judged based on the forecast-observation pairs (F_t, y_t) - Gneiting et al. (2007) suggest to base this assessment on the concepts calibration and sharpness, which are visualized in Figure 1.

In their words, calibration refers to the statistical consistency between the predictive distributions $(F_t)_{t=1,2,..}$ and the observations $(y_t)_{t=1,2,..}$ and is comprised of different modes - see Gneiting et al.

[4] We explicitly note here that a forecast is indexed by the forecast date t. This means that, when scoring forecasts, we also score by the forecast date and not by the date of the resolved target.

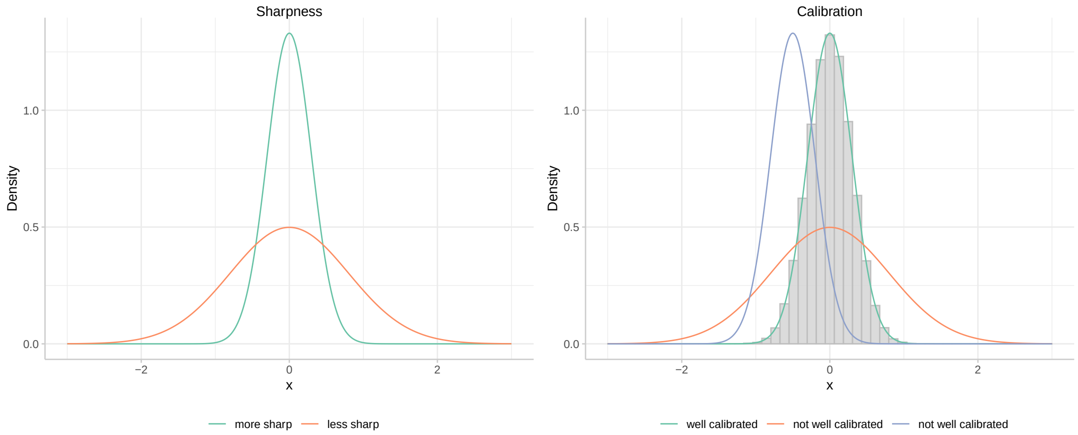


Figure 1: Illustration of the concepts *calibration* and *sharpness*, as described by Gneiting et al. (2007). Calibration refers to the statistical consistency between the predictive distributions (density plots) and the observed values of the series (grey histogram), while sharpness is a quality of the predictive distributions only. The general guidance is that forecasters should maximize sharpness of the predictive distributions subject to calibration.

(2007) for a full characterization of the concept. Thus, given a sufficiently large sample size, one can assess whether the data as it was observed was generally in line with the predictive distributions of the forecast model. Considering the right plot of Figure 1, the model represented by the blue distribution is systematically downward biased, as its probability mass is generally below the empirical distribution of the data. Conversely, the model represented by the orange distribution is not systematically biased, but overall under-confident. The model represented by the green distribution is well calibrated.

Sharpness is a feature of the predictive distribution (at time t) only and simply refers to the concentration of the predictive distribution (see left plot, Figure 1). The sharper a predictive distribution is, the more informative it will be to the user of a forecast. The goal is thus to devise a model that is overall in line with the observed data, but issues predictive distributions that are as sharp (and thus informative) as possible at any given time point.

In short, Gneiting et al. (2007) establish the result that the notion of an ideal forecaster is equivalent to the forecaster maximizing sharpness subject to calibration. In turn, this means that calibration is not a sufficient, but a necessary condition of an optimal forecast. When assessing forecast performance, both characteristics thus need to be judged and accounted for.

We will first introduce methods to assess calibration on its own and afterwards turn to proper scoring rules, which are convenient tools to summarize both sharpness and calibration in a single value.

3.2.2 Assessing calibration

The specific mode of calibration that focus usually lies on is probabilistic calibration (Bosse et al., 2022b). A forecast can be regarded as probabilistically calibrated if, for a given quantile level $\tau \in (0, 1)$, the proportion of observations falling below this quantile over time is also equal to τ. A natural and straightforward way to assess probabilistic calibration for quantile-based forecasts is thus by assessing the rates by which the observations are in line with the predictive distribution's quantiles, both in terms of the quantiles themselves and the central prediction intervals they induce. We thereby analyze the predictive distribution's *coverage*, described in more detail in the following.

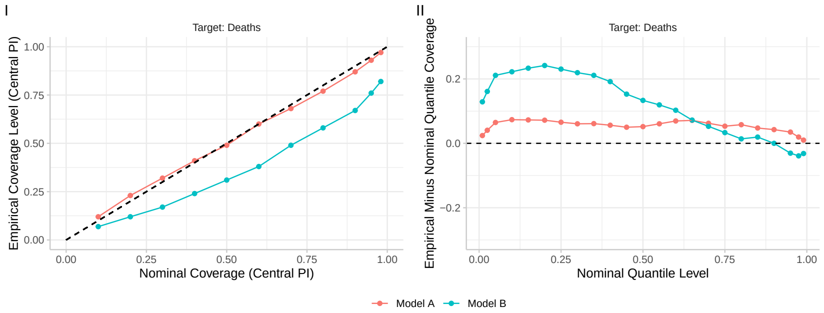


Figure 2: Central interval (I) and quantile coverage (II) rates of two component forecasts in the European COVID-19 Forecast Hub. Coverage rates can be used to assess a forecast distribution's probabilistic calibration.

Coverage

Interval coverage measures the proportion of observations that fall into a central prediction interval of the distribution. For instance, for the central 50% prediction interval, all observations that fell within the predictive distribution's 25% and 75% quantiles at the respective points in time are counted and divided by the total number of observations - the goal is to be as close as possible to the nominal coverage rate of 50%. Generally, a predictive distribution that is too wide and thus covers too many observations is called under-confident or conservative, while one that is too narrow is conversely called over-confident (Bosse et al., 2022b). Panel (I) of Figure 2 exemplary shows coverage rates for two models: model A's empirical interval coverage generally stays close to nominal coverage rates as indicated by the diagonal line, while model B's empirical interval coverage is generally too low, making the model over-confident. Within the literature around forecasting COVID-19, it is common to assess empirical interval coverage at the 50% and 90% or 95% levels (see for instance Sherratt et al. (2022a), Bosse et al. (2021)), to assess both the central tendency as well as the tail behavior of the predictive distributions.

Quantile coverage directly measures the proportion of observations that fall below a given quantile of the predictive distribution. It can thus convey more information than interval coverage, which does not distinguish between the upper and lower quantile of the central prediction interval (Bosse et al., 2022b). For instance, a predictive distribution could exhibit good performance at the lower tail end of the distribution, but not at the upper end, which quantile coverage can more directly diagnose. For instance, panel (II) of Figure 2 shows that model B's lower quantiles are generally set too high and upper quantiles are more in line with the observed data, resulting in overall too narrow prediction intervals as well as upward bias. Model A also shows some light upward bias that is more uniformly spread across the prediction quantiles.

Due to the nature of assessing rates by which observations fall into specific intervals, we generally need a sufficiently large sample to reliably assess coverage. This also means that estimates become less stable at the tail ends of the distribution, that is, for small α or small/large τ. For instance, consider the case where $\alpha = 0.05$ and the length of the series we are considering is $T = 60$: we are then basing our estimation of coverage on a presumably small number of data points that are not covered by this rather wide prediction interval, leading to a potentially noisy estimate. In particular, while we are ideally expecting 3 observations to fall outside the interval, it is entirely conceivable that even a well calibrated model could have e.g. 6 deviating observations due to natural variation - this model would then however only show a 90% coverage rate of the 95% prediction interval.

Bias

To solely assess whether a forecasting model is biased, that is, systematically over- or underpredicts the target of interest, one can utilize the bias metric as proposed by Funk et al. (2019). For integer-valued forecasts, they suggest the following metric:

$$B(F_t, y_t) = 1 - (F_t(y_t) + F_t(y_t + 1)).$$

In terms of bias, the ideal forecast would have exactly half its probability mass above and below the true value y_t, respectively. If more probability mass lies below the true value than above it, bias is negative - the extreme case occurs when the entire probability mass lies below true value, where we get $B_t = -1$. The opposite applies in the case where more probability mass lies above the true value.

Bosse et al. (2022b) extend the bias metric for quantile-based forecasts as follows. If the true value y_t is below the predicted median forecast $q_{t,0.5}$, the bias is

$$B(F_t, y_t) = 1 - 2 \cdot \max\{\tau | q_{t,\tau} \leq y_t\}.$$

Similarly, if the true value y_t is above the median forecast $q_{t,0.5}$, it is

$$B(F_t, y_t) = 1 - 2 \cdot \min\{\tau | q_{t,\tau} \geq y_t\}.$$

This can be interpreted as twice the difference between the quantile level that would ideally be closest to the observed data y_t ($\tau = 0.5$, i.e. the median) and the quantile level that is actually most in line with it. Concretely, if we interpret the quantiles as the endpoints of (central) prediction intervals, this is the outermost quantile of the predictive distribution such that the interval still contains the observed value y_t. As above, if the entire predictive mass is above or below the observed value, bias should take on the values 1 and -1, respectively. This is achieved by simply setting $q_{t,0} = 0$ and $q_{t,1} = \infty$.

Thus, if the observed value directly coincides with the median prediction $q_{t,0.5}$, bias is zero, otherwise, the forecast receives a penalty - the further the central mass of the predictive distribution is from the observed value, the larger the bias. The bias metric thus measures a forecast's general tendency to relatively over- or underpredict the target (Funk et al., 2019).

Thus, an important feature of the bias metric is that it is bound to the interval $[-1, 1]$ and thereby scores forecasts on a relative rather than an absolute scale - this means that we can directly compare different targets, even if their value ranges differ substantially.

3.2.3 Proper scoring rules

Suppose that y_t is the realization of a random variable under the true data-generating distribution G_t. As stated, a probabilistic forecast is issued as a predictive probability distribution F_t for this random variable. Scoring rules are then summary measures that assess the quality of the predictive distribution, by assigning a numerical score $s(F_t, y_t)$ based on the empirically observed agreement of the pair (F_t, y_t) (Gneiting and Raftery, 2007).

Further, denote $s(F_t, G_t)$ for the expectation of $\mathrm{E}[s(F_t, y_t)]$ under G_t. We then say that scoring rule s is *proper*, if

$$s(G_t, G_t) \leq s(F_t, G_t).$$

This means that the scoring rule is minimized if the true data-generating distribution is issued as the forecast distribution. Likewise, the scoring rule s is termed *strictly proper* for the strict inequality.
A (strictly) proper scoring rule thus incentivizes the forecaster to issue their true belief for the predictive probability distribution (Gneiting and Raftery, 2007). Consequently, propriety is a vital characteristic of scoring rules - if a scoring rule for a probabilistic forecasts were not proper, it could, for instance, incentivize a forecaster to report a more confident estimate than that which they actually believe in (Thorarinsdottir et al., 2013). This type of behavior is undesirable, as a potential forecast user wishing to be prepared for, say, an 80% range of plausible values for the outcome, would be underprepared as a result of such a forecast, with similar arguments applying for under-confident estimates. Recall the forecasting paradigm - proper scoring rules address calibration as well as sharpness, thus giving the forecaster the incentive to issue the sharpest predictive distribution that is expected to still be in line with the data (Gneiting et al., 2007).
In practice, we evaluate a model based on a series of forecast distributions it issued $(F_t)_{t=1,\dots,T}$, against the data as it was observed $(y_t)_{t=1,\dots,T}$. The series of scores can then be evaluated via their sum or, more commonly, their average

$$\bar{s}_F = \frac{1}{T}\sum_{i=1}^{T} s(F_t, y_t).$$

This allows us to compare the model based on F with a competing model based on H and consequently obtain a ranking between a set of forecasting models - as scoring rules are generally negatively oriented (lower values correspond to predictions that are more in line with the observations), the model with the lowest average score will lead such a ranking. This type of comparison is possible as aggregating scores via their sum or average preserves propriety of the scores (Bracher et al., 2021a).
We also note here that the specific scoring rule or metric used in an application is not a meaningless choice: even though we will mostly rely on the weighted interval score introduced in the next section, different scoring rules or other evaluation metrics can in principle induce different rankings of forecasts. In practice, it is thus wise to consider stakeholders' or decision makers preferences for forecast behavior, which can help guide the choice of assessment method.

3.2.4 Weighted Interval Score

Here, we introduce the weighted interval score (WIS), which is the main scoring rule used within this thesis (Bracher et al., 2021a). It is designed for use on probabilistic forecasts F that are issued as a set of discrete central prediction intervals, each with nominal coverage level α - or, equivalently, as a set of symmetric predictive quantiles q_τ which directly translate to central prediction intervals.
Each central prediction interval can be scored via the interval score (Gneiting and Raftery, 2007)

$$IS_\alpha(F, y) = (u - l) + \frac{2}{\alpha}(l - y)\mathbb{1}(y < l) + \frac{2}{\alpha}(y - u)\mathbb{1}(y > u), \tag{3.1}$$

where $\mathbb{1}()$ is the indicator function, returning 1 if the condition inside the parentheses is fulfilled and 0 otherwise. The three summands each have an intuitive interpretation. The second and third summands express under- and overprediction, respectively. The second summand assigns a penalty if the true observed quantity y falls below the lower endpoint l of the prediction interval - the size of the penalty is simply the absolute distance of the observation and the lower endpoint. The third summand then analogously assigns a penalty if y falls above the upper endpoint u. Both of these penalties are furthermore scaled by the nominal coverage level: a smaller α, which corresponds to a higher nominal coverage rate, induces a higher penalty if y does fall outside one of the endpoints (Bracher et al., 2021a).

Lastly, the first summand $(u - l)$ expresses the width of the central prediction interval and thus the sharpness of the predictive distribution F - if this term didn't exist, it would make sense to simply issue very large prediction intervals that are highly likely to contain the true observation y. It can thus straightforwardly be seen that this scoring rule addresses both calibration and sharpness.

Bracher et al. (2021a) extend this score for use on a predictive distribution F that consists of a set of such intervals, each with unique coverage level α. The set of interval scores is gathered and aggregated into the weighted interval score

$$WIS_{\alpha_{0:K}}(F, y) = \frac{1}{K + 1/2} \left(w_0 \cdot |y - m| + \sum_{k=1}^{K} (w_k \cdot IS_{\alpha_k}(F, y)) \right), \tag{3.2}$$

where usually the quantile weights are set to $w_k = \frac{\alpha_k}{2}$, and the median weight to $w_0 = \frac{1}{2}$.

It can be shown that the WIS is a discrete approximation of the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007), a well-known scoring function that measures the distance between the predictive and true distribution

$$CRPS(F, y) = \int_{-\infty}^{\infty} \left(F(x) - \mathbb{1}(x \geq y)\right)^2 dx.$$

Due to the format of the predictive distributions evaluated in this work (higher resolution around the tails), the WIS gives slightly larger weight to intervals with larger nominal coverage, in contrast to the CRPS (Bracher et al., 2021a). Furthermore, the WIS can be seen as a generalization of the absolute error (AE) for probabilistic forecasts. For a point-based prediction x_t, the AE is defined as

$$\text{AE}(x_t, y_t) = |y_t - x_t|.$$

It thus conveys less information than the WIS, but can be used when particular interest lies with the central tendency of the predictive distribution. When we are evaluating the AE, we are thus evaluating the predicted median value of a forecast, $q_{t,0.5}$.

All in all, the WIS is a parsimonious way to score forecasts that come in the shape of a set of discrete symmetric quantiles, and it is the most widely used scoring rule within the literature around forecasting COVID-19 (see for instance Sherratt et al. (2022b), Ray et al. (2022)). However, an important characteristic of the WIS is that it is not standardized and is thus valued on the natural scale of the data, similarly to the AE. This can be easily seen in equation 3.1, where the absolute differences of the observed value and the predicted quantile directly enter into the score. Thus, scores will typically increase if the level of the predicted series increases, as this is usually associated with larger absolute deviations from the target. This also means that the absolute values for over- and underprediction cannot be directly translated into an assessment of systematic bias, as there are no "ideal" absolute values for these components (Bosse et al., 2022b). To give an example, a forecast might overshoot a particular target during high incidence times, leading to a high absolute penalty in terms of overprediction, which will dominate its overall score - this however might mask a potential tendency of that same forecast to normally underpredict. This tendency would however be reflected in the aforementioned bias metric. It thus depends on the application and the evaluater's preferences whether interest lies more in over- and underprediction in absolute terms or in an overall relative tendency to over- or underpredict (Bosse et al., 2022b).

The absolute nature of the WIS can thus make forecast comparisons challenging - particularly in a case where there are missing forecasts for a model, thus leading us to the next point.

3.2.5 Comparisons of forecasts

An issue that often arises when aiming to compare different forecasting models is a potentially non-overlapping base of targets the models issued predictions for. As most scoring rules are not normalized and scale with the data, one model might look better than another in terms of simple average scores if it only predicted in periods that saw low incidence or that were otherwise comparatively ”easy” to forecast (Cramer et al., 2022). This would thus disincentivize forecasters to predict in periods that they perceive to be more challenging - this is especially undesirable because these periods (for instance, changes in trends, high levels of incidence) are often of special interest to decision makers (Reich et al., 2021).

One can address this by computing a relative score that is based on pairwise comparisons of models, as developed in Cramer et al. (2022). For a pair of models denoted l and m, first a measure of relative skill is computed

$$\theta_{l,m} = \frac{\bar{s}_l}{\bar{s}_m},$$

where $\bar{s}_l$ and $\bar{s}_m$ denote the average scores the models achieved on the targets both models predicted on. In principle, one can base these comparison on any scoring rule that gives either exclusively positive or exclusively negative values, but we follow the relevant literature and thus use the WIS.

For each model l, one then computes

$$\theta_{l\cdot} = \left(\prod_{m=1}^{M} \theta_{l,m} \right)^{\frac{1}{M}},$$

to obtain a relative score of model l with respect to all other available models. The choice of the geometric mean as an aggregation function over models is convention in the literature, but Ray et al. (2022) state in that the choice of aggregation function has not shown itself to be critical in their application, and that one could thus also simply use an arithmetic mean.

This relative score can be interpreted as a performance measure of model l with respect to a model with “average” performance. If interest lies in a direct pairwise comparison with a specific model m, one can instead consider the ratio of these relative scores

$$\phi_{l,m} = \frac{\theta_{l\cdot}}{\theta_{m\cdot}}. \tag{3.3}$$

Calculating this ratio for all model pairs that are of interest results in a “pairwise tournament” for all models in the set - this approach is implemented in the `scoringutils` package (Bosse et al., 2022a). For negatively oriented scoring rules, the ratio will be smaller than 1 if model l outperformed model m on their set of shared targets and larger than 1 if it did not. Note that this mode of pairwise comparison still requires the assumption that it is equally hard to perform relatively well to other models at all forecast dates (Cramer et al., 2022).

If one is interested in concisely summarizing the skill of single models rather than performing comparisons between all pairs of models, one can choose a baseline model's B relative score $\theta_{B\cdot}$ as the denominator. When using the WIS as the scoring rule, this results in the measure that is commonly referred to as ”relative WIS” (Cramer et al., 2022). Analogously to above, a ratio below 1 corresponds to a model overall outperforming the baseline model, while a score above 1 means that the model did not succeed

in clearing baseline performance.
In section 5.1, we use the method of pairwise comparisons to compare groups of models, more specifically the aforementioned modeling strategies from section 2.1, against each other. The average scores are thus calculated at the level of model groups. To not skew results by the differing numbers of models available for a group over time, scores for a given group are first averaged by forecast dates, before computing an average over (a subset of) the study period.

This method of comparing component forecasts was largely developed to address missingness in component forecasts and can thus be used to compare a larger number of models with differing amount of availability. When interest instead lies in directly comparing the performance of two specific models against each other, especially when one model is a more established "benchmark" model that a newly proposed model should be compared against, it can be more straightforward and natural to directly compare the average scores the two models obtained over all considered forecast dates and locations. This is also the approach taken by e.g. Bosse et al. (2021).
Importantly, while this style of forecast comparison addresses the problem of non-overlapping models, it is still based on average scores and is thus most heavily influenced by targets with large nominal scores.

3.2.6 Model distance

We so far discussed scoring rules, which, as stated, are used to assess the pairs (F_t, y_t). In contrast to this, divergence functions d are used to assess distance between a pair of distribution functions (F_t, H_t) (Thorarinsdottir et al., 2013). These can be used to score the predictive distribution F_t against an estimate of the true data generating distribution $\hat{G}_t$, but also to measure the distance between two competing predictive distributions F_t and H_t.
As it is fundamentally a measure of distance, the divergence function needs to fulfill $d(F, F) = 0$ and $d(F, H) \geq 0$ for all F and H. Furthermore, as they can be used to score a predictive distribution, it also needs to fulfill the notion of propriety as defined previously. Analogously to before, this means that in expectation, the true data-generating distribution will minimize the expectation of the divergence function when evaluated against the estimated empirical distribution $\hat{G}_t$.
Within the context of this work, we only use divergence functions to measure the distance between different predictive distributions. The precise distance function we use towards this end is the 2-order Cramer distance or integrated quadratic distance (Thorarinsdottir et al., 2013). It is defined as the integral

$$\mathrm{CD}(F, H) = \int_{-\infty}^{\infty} (F(t) - H(t))^2 \, dt.$$

Incidentally, this is the divergence function associated with the CRPS as defined previously - this means that if we evaluate F_t against a single observation instead of another distribution, it will reduce to the CRPS (Gneiting and Raftery, 2007).
Wang et al. (2022) developed an approximation of the Cramer distance for (potentially unequally-spaced) quantile-based forecasts as follows:

$$\mathrm{CD}(F, H) \approx \sum_{i=1}^{2K-1} (a_i^F - a_i^H)^2 (q_{i+1}^P - q_i^P), \tag{3.4}$$

where q^P contains the $2K$ pooled and ordered predictive quantiles[5] from F and H and a_i^F is a vector that, for each quantile q_i in the pooled quantile vector, contains the minimum quantile level τ for which the associated quantile q_τ^F is larger than it. We write this condition as

$$a_i^F = \min\{\tau | q_\tau^F \geq q_i^P\},$$

with an analogous definition for H. Similarly to the scoring rules discussed above, the absolute difference in quantiles directly enters into the function in equation 3.4. Thus, the Cramer distance also scales with the data, meaning that we will naturally see larger measures of distance during periods with higher incidence levels.

3.3 Ensembling Methods

In this section, we introduce different methods that can be used to combine a set of quantile-based component forecasts into a single (also quantile-based) ensemble forecast. Generally, the most widespread methods for combining forecasts of this format are those that use a quantile-wise mean or median, with the additional option to more heavily place weights on certain component forecasts.

The (weighted) quantile-wise mean and median are also the aggregation methods that we use within this thesis, but note that other techniques are also possible. For instance, one could first derive a continuous probability distribution for each quantile-based forecast, and subsequently average these, before again taking discrete quantiles from the combined predictive distribution. To our knowledge, these techniques were experimentally tried for application in the U.S. Hub, without promising gains in performance. Additionally, it is possible to combine the mean and median approach with trimming methods, which - broadly speaking - remove a pre-determined number of most extreme valued forecasts before combining the remaining ones. An extensive evaluation of these methods was performed in Taylor and Taylor (2021), also without clear gains in performance.

In this section, we thus want to formally introduce the different aggregation techniques available to combine quantile-based forecasts, as well as methods that can used to estimate weights for component forecasts based on their recent performance.[6] For the mean ensemble, each quantile of the ensemble forecast is simply computed as the mean of that quantile from the component forecasts, that is

$$q_{E,t,h,\tau} = \sum_{m=1}^{M_t} w_{m,t} \cdot q_{m,t,h,\tau}.$$

For an equally weighted mean ensemble, the weights are simply set to $w_{m,t} = \frac{1}{M_t}$ for all models, where M_t is the number of models predicting for the given location and target series at time t.

We define the weighted median as

$$q_{E,t,h,\tau} = \min\left\{ q_{m^*,t,h,\tau} \in Q_{t,h,\tau}^P : \sum_{m=1}^{M_t} w_{m,t} \cdot \mathbb{1}(q_{m,t,h,\tau} < q_{m^*,t,h,\tau}) = 0.5 \right\},$$

where $Q_{t,h,\tau}^P$ denotes the pool of quantiles for level τ at time t and horizon h from all available component forecasters. The weighted median is thus the value for which the associated weight of all values smaller

[5] With a slight abuse of notation as the subscript i in this case indexes the position in the vector q^P, whereas usually the subscript refers to the quantile level τ. However, since $\tau \in (0,1)$, and is thus strictly smaller than 1, and $i \in \{1, 2, ..., 2K\}$, and is thus larger than 1, we believe the distinction is clear enough.

[6] The notation used within this section was to a certain degree inspired by Ray et al. (2022).

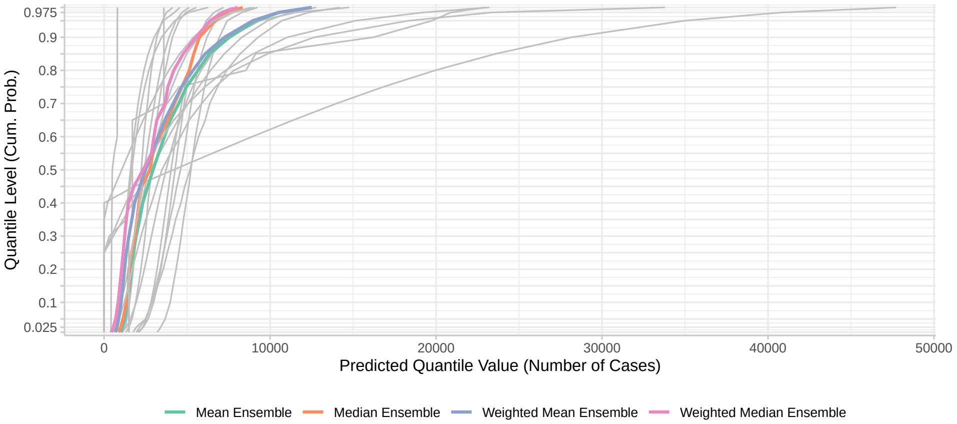


Figure 3: Four ensembles resulting from a set of 20 component forecasts that issue forecasts for weekly incident cases in Germany on July 5, 2021 (Monday). The depicted forecasts are made for a two-week forecast horizon, with target date July 17, 2021 (Saturday). Note that forecast distributions are issued as a set of 23 non-equally spaced discrete quantiles (each one indicated by a tick on the y-axis), the depicted curves are thus obtained through linear interpolations of the quantiles. Component forecasts' predictive distributions are shown in grey, while the different ensemble forecasts are colored according to the legend. Ensemble forecasts are obtained by taking a (weighted) quantile-wise mean or median of component forecasts' quantiles - visually, for each quantile level on the y-axis, the component quantiles are horizontally combined. The weights for the weighted mean or median ensemble were obtained by inverse score weighting. Plot inspired by Taylor and Taylor (2021).

than it is half of the total weight, for which we have the common restriction $\sum_m^{M_t} w_m = 1$. If this value is not exactly attained within the pool of quantiles, linear interpolation is performed. For our purposes, we used the `weighted.median` function from the `spatstats` package (Baddeley and Turner, 2005). Similarly to above, if we set $w_m = \frac{1}{M_t}$ for all models, we simply obtain the value which has half of all other values lying below it, that is, the regular median.

In Figure 3, we show an exemplary illustration from our dataset of how the respective ensemble forecasts are formed, specifically for the incident case series in Germany, in this case from a set of 20 component forecasts. Especially at the upper end, predictive quantiles are fairly spread out - while most forecasters place the uppermost quantile ($\tau = 0.99$) to a value below 15000, some issue substantially more wide predictive distributions, with the most extreme value just below 50000. We can observe that the combination methods are differently impacted by these outlying forecasts, with ensembles based on the mean issuing larger forecasts for the upper quantile levels. The bigger resistance to outlying forecasts is a central feature of the median ensemble (Sherratt et al., 2022a). A potential downside of the median ensemble is that its quantiles can collapse, especially if it is based on a smaller number of component forecasts (Bracher et al., 2021a).

To determine weights for ensemble forecasts, several strategies have been proposed and applied in the literature, particularly around ensemble forecasts in the Hub papers. We mention those which we will also utilize within this work.

A straightforward way to weight forecasts is by setting weights inversely proportional to recent scores obtained by component forecasts, to our knowledge first applied in the literature around forecasting COVID-19 by Bracher et al. (2021b). The resulting weighting scheme is thus termed inverse score or inverse WIS weighting, with weights at time t simply set to

$$w_{m,t} = \frac{1}{\bar{s}_{m,t}},$$

where $\bar{s}_{m,t}$ is the mean WIS value obtained by model m over a rolling window of recently realized targets. It is thus computed as a simple moving average, that is

$$\bar{s}_{m,t} = \sum_{r=t-1}^{t-a} \sum_{h=1}^{\max\{h:r+h\leq t\}} \text{WIS}(q_{m,r,h,1:K},\ y_r).$$

Note thus that targets that have not yet realized by the current forecast origin are excluded from the estimation. In a real-time setting, this is of course automatically enforced, while in a retrospective analysis such as here, combinations of forecast dates and horizons that lie in the future relative to the forecast origin need to be actively excluded. The weights are subsequently normalized such that they sum to one. Other scoring rules beside the WIS are in theory also possible, although we are not aware of any applications - the WIS remains the natural choice, as it is the most parsimonious way to score forecasts in terms of both calibration and sharpness.

Regarding the possibility of models having missed forecasts within the window, different courses of actions are possible. These forecasts could be excluded from the weighting scheme entirely, which has the disadvantage of reducing the model set for the resulting ensemble. Taylor and Taylor (2021) straightforwardly compute the average score based on only the available historical forecasts, but this can be problematic due to potentially favoring models that (perhaps even strategically) did not issue forecasts for targets that were perceived as difficult and consequently have a higher probability to induce larger scores. Alternatively, missing scores can be imputed, with Bracher et al. (2021b) suggesting to impute a missing score by the worst score achieved by any model for that particular target. This can be regarded as a pragmatic approach to include all available models in the weighting scheme, without the risk to favor forecasts that shirked forecasting in more difficult weeks, although it could potentially also skew weights in cases where outlying forecasts with large absolute scores exist within the relevant set of realized forecasts. Yet another possibility is to base estimation of weights on the relative WIS as described in section 3.2.5, as is done in Ray et al. (2022).

The window size a is a parameter that effectively governs the degree to which more distant historical forecasts impact the weight estimation. The more historical forecasts are considered, the smaller the variance of the weight estimation, while bias will reduce if we exclusively base weight estimation on more recently made forecasts (if we assume forecast performance to be non-stationary). Note also that the smaller the window size, the more weight is proportionally placed on forecasts that were made for shorter time horizons - for instance, if $a = 3$, only one three-week ahead and no four-week ahead forecast will have realized by the forecast origin. We would argue that this is somewhat desirable, as scores at longer horizons usually induce larger absolute scores and thus influence the weight estimation more. For larger windows, this effect somewhat disappears, as the number of forecasts considered for each horizon are relatively more similar.

Due to the fact that performance of component models generally is non-stationary and more distant scores can thus be regarded as having less predictive power for current forecast performance, we additionally wanted to trial an ex-ante weighting scheme applied to the scores themselves. This scheme should more heavily weight recent scores, while still including more distant scores, albeit to a lesser degree. A natural choice is thus an exponential smoothing approach, thus changing the computation to

$$\bar{s}^{e}_{m,t} = \sum_{r=t-1}^{t-a} \alpha \cdot (1-\alpha)^{(t-r-1)} \sum_{h=1}^{\max\{h:r+h\leq t\}} \text{WIS}(q_{m,r,h,\tau},\ y_r).$$

As we still use a rolling window and thus don't include all historical forecasts, we additionally scale the exponential smoothing weights such that they sum to one. While this nature of essentially cutting off observations after a certain lag is to our knowledge not usual practice for exponential smoothing approaches, we regard this as unproblematic. This is due to the fact that more distant forecast, due to the nature of the exponential decrease, would receive lower weights in any case. More importantly, this practice saves some trouble that can potentially arise with imputing missing forecasts.
Another possibility for weighting component forecasts is the quantile regression average (QRA) ensemble, as introduced by Nowotarski and Weron (2015). It is implemented in the `quantgen` package (Tibshirani and Brooks, 2020) and chooses weights such that they directly minimize the historical weighted interval score. For this approach, weights are chosen according to the following objective function

$$\min_{\boldsymbol{w}} \sum_{k=1}^{K} \sum_{r=t-1}^{t-a} \psi_{\tau_k}(y_r - \sum_{m=1}^{M} w_m q_{r,m,k}), \tag{3.5}$$

again with the restriction that the elements of the weight vector $\boldsymbol{w}$ must sum to one. ψ_τ is the "pinball" loss for quantile level $\tau \in (0,1)$, that is

$$\psi_\tau(v) = \max\{\tau v, (\tau - 1)v\}.$$

As can be seen in Equation (3.5), the weight estimation method is based on minimizing scores by taking a weighted mean of the forecast distributions' quantiles. The natural choice is to consequently also use the resulting weights w_m in a mean ensemble.
Brooks et al. (2020) found the resulting ensemble to be competitive to an equally weighted median ensemble, for forecasting COVID-19 deaths during the year 2020. An issue with this method as it is implemented is that it needs full data for all component forecasts within the relevant window. As weight estimation is based directly on the set of predictive quantiles, this makes imputation of missing values more challenging, contrary to the weighting methods based only on forecasts' scores. Furthermore, as the pinball loss scales with the target value, the method is generally vulnerable to be influenced by locations or single targets with larger absolute levels. However, the same thing can be said for the inverse score weighted methods.

For all weighting methods, it is generally possible to allow for more flexibility and separately estimate weights for the different quantiles, as well as the forecast horizons. Ray et al. (2022) state that they did not observe any consistent improvements from such approaches, and accordingly resorted to simpler weighting schemes. In fact, additional issues could arise from this. For instance, a desirable property of forecast distributions is that they have a widening cone of uncertainty with increasing forecast horizon, to account for the fact that uncertainty is greater for targets that lie farther in the future. This is generally also a feature that is upheld across single component forecasts - when different forecasts are however weighted unequally across horizons, this continuity could be disturbed. Furthermore, when allowing flexible weights across quantiles, one needs to incorporate noncrossing constraints.
For our study, we did however decide to generally estimate weights separately by location. We do this to counteract the effect of locations with higher observed values influencing the estimated weights too heavily, as well as to pragmatically account for the fact that the set of models varies considerably by location. Furthermore, weights are computed separately for the two target series (cases and deaths), which is common practice.

3.4 Statistical analysis of scores

The assessment of forecasts' scores conducted within this thesis, be it for the different modeling strategies or alternatively proposed ensemble methods, will largely not depend on statistical testing. Most of the works cited in this thesis that similarly perform retrospective forecast evaluations follow the same practice. There do exist methods to test the significance of differences in forecast scores/skill between models, the most common being the Diebold-Mariano test (Diebold and Mariano, 1995), which however lacks an extension that accounts for correlations between forecasts issued for different target series and/or locations (Bracher et al., 2021a).

To compare the performance of different forecasts and/or groups of models, one could also conceive of a regression analysis with scores as the dependent variable, which would have the added benefit of being able to systematically investigate effects and reduce the dimensionality of the results. Such an analysis could account for idiosyncrasies arising at certain locations or time points via e.g. the inclusion of random effects, but would fundamentally suffer from the issue that scores as they are recorded in the data are correlated across time and component forecasts, and furthermore generally have large outliers which can heavily influence results and thus weaken the trust in reported p-values. Log-transforming the scores to bring them more in line with a normal distribution and reduce the effect of outlying values is also not a viable alternative, as the log-transformation does not preserve the propriety of the scores.

We can thus view the analyses performed in this work as having a more descriptive and introspective nature - we feel that this approach is to a certain degree more honest than simply performing tests while being aware of these issues, the results of which could consequently not be trusted due to the aforementioned issues. We thereby follow the advice of Bosse et al. (2022b), who generally advise against formal statistical testing of forecasts “in most applied settings”.

We state this to explicitly show the limitations of our analysis and argue that some of the knowledge distilled in this work should be checked against another year's results, to increase trust in the results.

4 Data

In this section, we introduce the dataset that is used in this thesis. We first generally introduce the European COVID-19 Forecast Hub (Sherratt et al., 2022a), and then introduce the particular subset our work is based on.

4.1 European COVID-19 Forecast Hub

The data used in this thesis stem from the European COVID-19 Forecast Hub, thereafter referred to as the "European Hub" or just "Hub".[7] It was instigated by the ECDC in collaboration with the Epiforecasts team at the London School of Hygiene and Tropical Medicine (LSHTM) in March of 2021 to collate weekly forecasts for COVID-19 incident cases and deaths[8] for 32 European countries. Models were developed and forecasts subsequently issued by independent modeling teams. The Hub was modeled after a similar previous effort in the United States, the United States COVID-19 Forecast Hub (Cramer et al., 2021), thereafter referred to as the "U.S. Hub". Furthermore, the preceding German-Polish Forecast Hub (Bracher et al., 2020) was largely synchronized with the European Hub. The Hub's primary goal is stated to "provide reliable information about the near-term epidemiology of the COVID-19 pandemic to the research and policy communities and the general public" (Sherratt et al., 2022b).

In general, a modeling hub is a coordinated effort, in which one or more common prediction targets, as well as a common format for prediction, are agreed upon and centrally implemented (Reich et al., 2022). This serves the purpose of facilitating model evaluation and development by making model predictions comparable. Furthermore, it makes predictions suitable for aggregation, that is, for generating ensemble predictions. A central advantage of the "Hub"-methodology is thus the potential to synthesize results from different modeling approaches and teams. This can alleviate some problems that arise in less coordinated research efforts - for instance, Metcalf and Lessler (2017) remark that, following major disease outbreaks, there is often an explosion of modeling studies, but the usefulness of the thereby generated research is limited as quality of data and methods used vary, and there is often no follow-up for synthetization of results. The Hub format, in contrast to this, standardizes data quality[9] and facilitates evaluation and thus synthetization of results through a set of shared targets. The format has some precedent both in other fields, for instance climatology or ecology (Warszawski et al., 2014), as well as in epidemiology itself - some notable examples include forecasting influenza in the United States (Reich et al., 2019) as well as dengue fever in Puerto Rico and Peru (Johansson et al., 2019).

The common prediction targets for the COVID-19 Hubs are weekly COVID-19 case and death counts in the respective locations (U.S. states, European countries). Later, hospitalization rates were also added as a target series to both Hubs, but we don't consider them within the scope of this work.
There were no restrictions for participation in the Hub, meaning that in theory anyone could participate, although most submissions did come from academic modeling teams. Participating teams or individuals were free to only submit forecasts for any subset of combinations of locations and target types - thus, there were both models that only predict for certain locations and/or for only cases or deaths. Forecast

[7] The information in this section regarding the European Hub's operational/methodological approaches, unless otherwise indicated, is taken from the Wiki at https://github.com/covid19-forecast-hub-europe

[8] That is, the number of newly confirmed cases and deaths within a given week. Other forecasting efforts might additionally or alternatively ask for predictions of *cumulative* counts.

[9] To a certain degree: In case of the European Hub, teams are free to use whatever data sources they wish, but are told that forecasts will be evaluated against the publicly available JHU data and are thus recommended to base their methods on the same data source. Furthermore, additional data sources for e.g. vaccination and testing data are recommended.

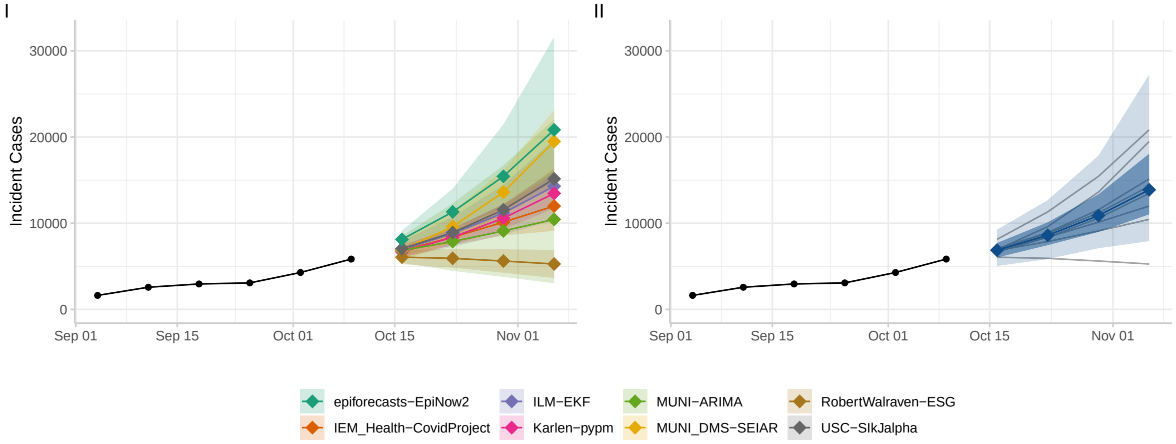


Figure 4: Example trajectory with forecasts for incident cases from the Czech Republic. Following the last available truth data point on October 9, 2021, modeling teams submit forecasts for weekly incident cases, for one to four weeks into the future, respectively as a predictive distribution via a set of discrete quantiles. Shown is the series as it realized until October 9 (black line), and the trajectories of the respective median predictions for one to four weeks into the future, with 50% central prediction intervals (shaded areas). Panel (I) shows the predictions of the component forecasts, panel (II) the resulting aggregated prediction of the official European Hub ensemble, with 50% and 90% central prediction interval (shaded areas) and median predictions by component models (grey lines). Plot inspired by Brooks et al. (2020).

dates were standardized and based on the Epidemiological Week (EW) format as defined by the U.S. CDC, where each week starts on a Sunday and ends on the following Saturday. Forecasts were submitted by Monday and thus scored against the truth data as it realized on the Saturday of the same week for 1-week ahead forecasts, and accordingly by the following Saturdays for larger forecast horizons.

The truth data source that models are evaluated against stem from JHU (Dong et al., 2020), which collate and make available daily cumulative counts of COVID-19 cases and deaths, for a large set of countries around the world. Since the Hub asks for weekly forecasts of incidence, the truth data is obtained by taking weekly differences of JHU national data. This can in theory be problematic, as data are subject to revision and negative values for incident counts are thus possible.

As stated before, the modeling team's forecasts are issued in a probabilistic manner, more specifically as a set of 23 discrete quantiles of the predictive distribution, at non-equally-spaced levels between 0.01 and 0.99. Consider Figure 4 for an illustration of the prediction format, exemplary for case numbers in the Czech Republic with the forecast date October 11. Panel (I) shows the individual predictions made by the eight component forecasts that participated on this date, with median predictions as well as the central 50% prediction intervals. We can see that most forecasters predicted a continued rise in case numbers, albeit with varying levels of certainty. Panel (II) shows the corresponding forecast made by the official Hub ensemble (additionally with the 90% prediction interval), which consequently also predicts a continued rise in case numbers.

For both the European and the U.S. Hub, the central communication tool lies with the Hub ensemble, which was generally found to yield the most consistent performance for both target series, compared to the performance of single component forecasts (Sherratt et al. (2022a), Cramer et al. (2022)). Both Hubs initially used a mean ensemble, but switched to the median ensemble after observing that it was more robust to occasional outlying forecasts - subsequent evaluations also showed that the median outperformed the mean as a combination method for the ensemble (Sherratt et al. (2022a), Ray et al. (2022)). To be included in the official European Hub's ensemble, models had to provide a full set of 23 quantiles for all four forecast horizons, as well as pass a sequence of automated checks for their predictions, that is, that their quantile predictions were non-negative, integer-valued and did not cross (Sherratt et al., 2022a).

The Hub also includes a baseline model, which is the same as the one that is used in the U.S. Hub. For each forecast origin, its forecast for the median incidence for both target series is equal to the value that was observed in the most recent week. For uncertainty around the median, the other predictive quantiles are taken from Monte Carlo approximations of the empirical distribution function that is induced by the first differences observed in the respective series (Cramer et al., 2022). In essence, this model can thus be seen as a random walk model with non-stationary variance.

Including a baseline model in a collective forecasting effort serves several goals: First, including such a “naive” model serves the purpose of providing a sort of “minimum” performance threshold that models should be able to clear - being able to report that a model has higher skill than the baseline thus gives validity to the performance of that model. Second, given the fact that most prominent scoring rules scale with the target and thus never obtain a sort of “optimum”, reporting performance relative to the baseline is a pragmatic strategy to compare performance across dimensions.

However, we note here that while the introduction of a baseline aims to fulfill the purpose of standardizing model comparison, it is not without alternative and its specific form can impact model assessment (Sherratt et al., 2022b). For instance, if the baseline model systematically performs worse during some periods, this might portray individual model performance more favorably - see section 5.1 for an example. We also refer to Bracher et al. (2021b), who, in their analysis of the German-Polish Hub, retrospectively included alternative baseline models to compare individual model performance against.

An important difference between the European and U.S. Hub pertains to the number of forecasters participating, as well as to participation continuity. Specifically, while the U.S. sees an average of over 30 submissions per week, most of which submit forecasts for all or most of the 50 U.S. states (Ray et al., 2022), the specific forecasters in the respective model sets as well as the total number of available models substantially varies by location in the European Hub. Furthermore, entry and exit of forecasters (and thus less consistent records of past performance) are more of an issue in the European Hub. This comes with its challenges, which will become relevant in section 6.1.

Lastly, we discuss some results from previous Hub studies in relation to the differences between (ensemble) forecasting the two target series, cases and deaths. In general, it has been found that the ensemble's case forecasts become substantially less reliable beyond the one or two week horizon, while forecasts for deaths generally remain more reliable across all four forecast horizons, for both the U.S. and the European Hub (e.g. Reich et al. (2021), Sherratt et al. (2022a)). In fact, both Hubs officially recommend to focus on either the one or the one and two week horizons when considering case forecasts.[10] In particular, it has been noted that both component and ensemble forecasts struggle with predicting dynamic changes for cases, while they are more often able to predict changes in trends for deaths (Reich et al., 2021). Bracher et al. (2021b) (among others) argue that the higher predictability of the deaths series relates to the fact that, due to the larger time horizon between infection and death, they are a lagged indicator of the pandemic and thus are less affected by changes in e.g. interventions for the considered time horizons. Most notably, past reported cases can thus serve as a predictor for deaths, and there has been some evidence that models that use past cases as an input to forecasting deaths tend to perform better (Cramer et al. (2022), Bosse et al. (2021)).

[10] See https://covid19forecasthub.org/doc/ensemble/ and https://covid19forecasthub.eu/background.

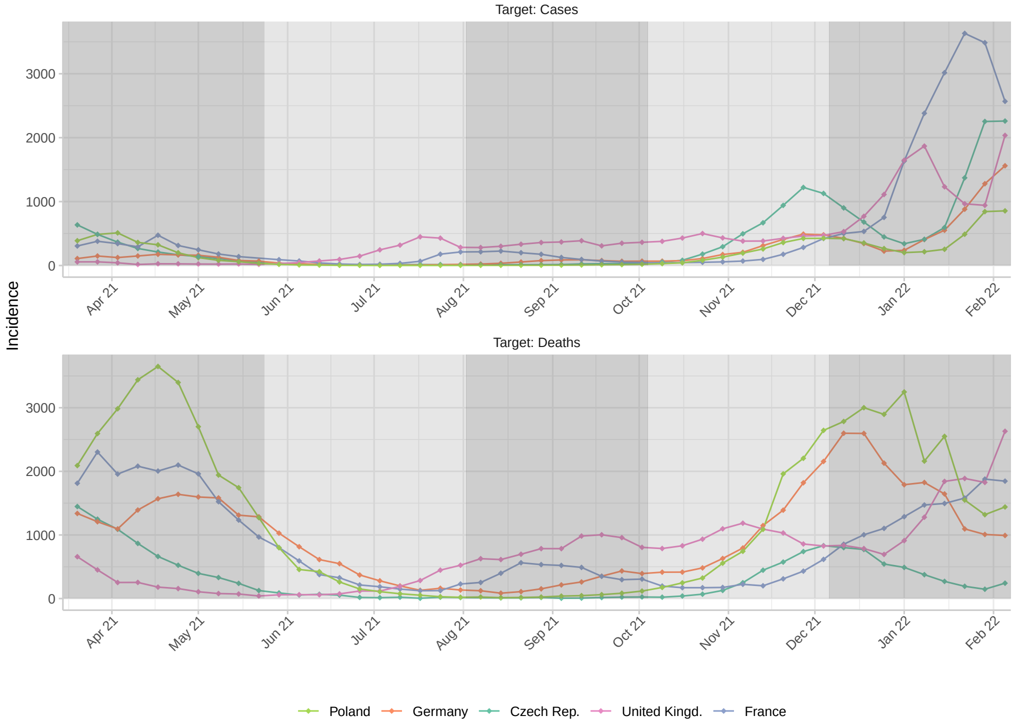


Figure 5: Observed trajectories of the cases and deaths series for the locations from the European COVID-19 Forecast Hub considered within this study. Note that in this plot the time series for cases are incident counts per 100000 inhabitants (and were thereby scaled by the respective location's population), to facilitate readability of all locations' observed trajectories. The time series for deaths are direct incident counts. The alternating gray background indicates a categorization of periods, based on slicing the time period into five equally sized sub-periods. The equality of the y-axes' limits is a coincidence.

4.2 Data subset

The subset of the European COVID-19 Forecast Hub that we base our analysis on comprises data from five European countries, with forecasts issued for both target series (cases and deaths) between March of 2021 and the end of January of 2022.

We restricted ourselves to the countries that had the largest number of available models within the Hub - specifically, these are the Czech Republic, France, Germany, Poland and the United Kingdom. This is due to the fact that most of the analyses we aimed to perform required a sufficiently large model base as support. We decided to only use data up until the end of January of 2022 (and thereby targets that realized up until the end of February 2022), since heading into the spring months of 2022, various regulations and in particular testing criteria changed for many locations. We thought that this could make the time periods less comparable. In total, this leaves us with a total of 47 weeks and thereby 47 forecast dates with fully resolved forecasts. Furthermore, we restricted our analyses to models that submitted the full set of 23 quantiles for each of the four horizons - otherwise, both evaluation of models as well as building ensembles would be significantly more challenging.

In Figure 5, we visualize the observed time series for the countries, separately for the two target series cases and deaths. We generally observe that while absolute values are often substantially different across

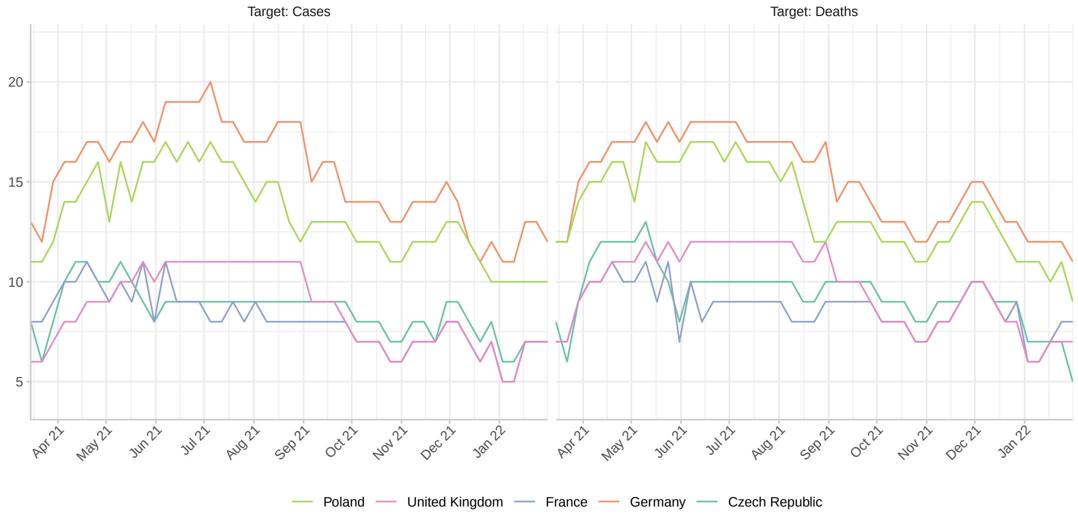


Figure 6: Overall availability of component forecasts submitted to the European COVID-19 Forecast Hub for the five locations and two target series considered in this study, visualized over time and separately for the two target series (incident cases and incident deaths). Note that wherever the line plot for France or the Czech Republic isn't visible, it is overlain by that of the United Kingdom, as the locations have the same number of models during these times. As these plots show the availability of component forecasts, the European Hub baseline and ensemble model are excluded from the total number.

countries and “waves” are not perfectly aligned, there nevertheless exist some common characteristics, such as generally lower incident counts during the summer months as well as rising incidence towards the winter months, for both target series. As we will sometimes be interested in comparing forecast performance over time, but found that individual forecast dates provided too high of a resolution, we followed Taylor and Taylor (2021) in dividing our study period into 10-week periods[11]. These are indicated in Figure 5 by the alternating gray background. As scores are often dominated by higher levels of the underlying target series, this categorization will also be helpful, as it allows us to separately assess performance during periods with lower levels of incidence.

For cases, incident counts are declining at the beginning of the study period, followed by mostly lower levels in subsequent periods, which are however also characterized by intermittent periods of rising levels. The series show rising trends approaching the winter months, followed by the overall highest levels towards the end of the study period. For deaths, we similarly observe overall declining levels at the start of the study period and rising levels towards the end, although differences in absolute levels are not as high. The U.K. (and to a certain degree France), which also has higher case numbers during the summer months, also experiences higher mortality during this time.

Recall our mention of the theoretically possible negative incident counts in the previous section. In our subset of the data, this occurred twice, once for the cases series in France and once for the deaths series in the Czech Republic - we excluded these observations and thus the forecasts that predicted for these particular targets.

We visualize total as well as individual model availability in Figures 6 and 7, respectively. Regarding the total number of available models by location, we observe that, after an initial period of uptake and recruiting, the modeling base is largest during the summer months and into the fall of 2021, after which it generally drops, with an intermittent slightly higher participation in December of 2021. Between coun-

[11] since we have 47 weeks in our studied sample, we actually divide into 2 10-week and 3 9-week periods

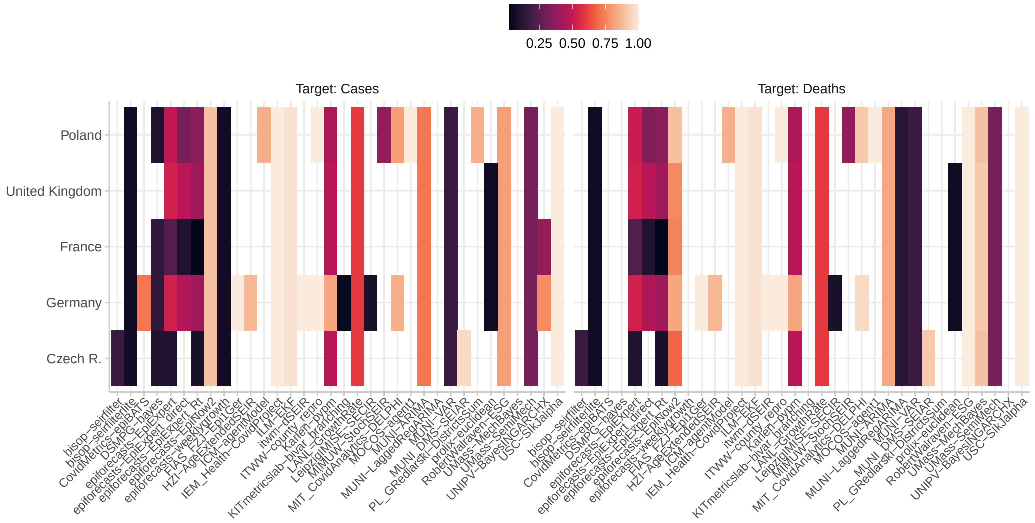


Figure 7: Availability of each model submitted to the European COVID-19 Forecast Hub for the five locations and two target series considered in this study. Availability was calculated as the proportion of the study period for which a model was available, for the given combination of location and target series. Missing panels refer to the case where a model never issued forecasts for a combination. Individual model availability generally varies, with only a small number of models participating for the entire period under study. Note that the European Hub baseline and official ensemble model are not shown, but do have perfect availability.

tries, we observe that Poland and in particular Germany generally see higher levels of participation, while the Czech Republic, France and the United Kingdom overall have similarly lower numbers of forecasters participating.

At the level of individual model availability (Figure 7), pronounced fluctuations are also observable. A small number of models is available for the entire study period and across all locations, such as `IEM_Health-CovidProject`, while others, such as `ITWW-country_repro`, submit forecasts for the entire study period, but only for a subset of locations. Usually, we however see "non-perfect" availability for one or most locations, which can result from different base behavior: usually, this is due to modelers either starting to participate at a later date and/or dropping out of the Hub entirely (such as `Karlen-pypm`), while others might only miss forecasting some weeks before subsequently participating again (such as `epiforecasts-EpiNow2`). In general, this varying model base can be an issue, for both forecast evaluation and making ensemble predictions. One can deal with these issues by, for instance, only taking a full forecast set. However, we deliberately wanted to deal with the data as-is and develop and conceive strategies that work with and around this characteristic of the data, as we felt that this was the most realistic and helpful for the situation at hand.

5 Model and Ensemble behavior

In this section, we perform some more exploratory analyses into model and ensemble performance within the European Forecast Hub. In the section following afterward, we consequently aim to explore whether some of the knowledge gained can be leveraged to establish methods that could in principle improve ensemble performance in real-time.

Specifically, we explore the performance of different modeling strategies, over the study period and the locations and target series we consider in this study. Furthermore, we investigate the effect of adding a model to an existing ensemble, to better understand ensemble performance with respect to its member models.

5.1 Modeling Strategies

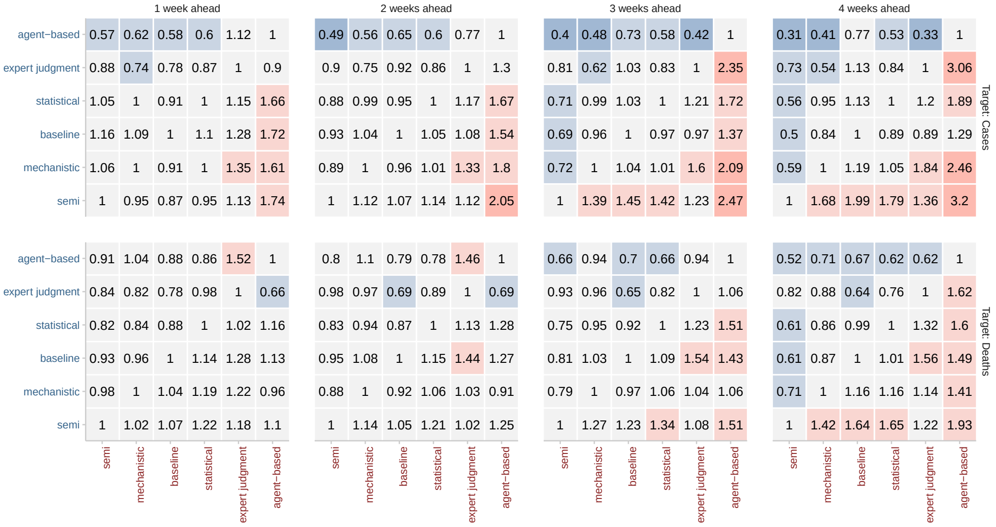
It remains an open topic of research which modeling strategy is best suited for epidemiological forecasting - and in particular, whether it is necessary to explicitly include mechanistic assumptions to yield accurate short-term forecasts (Funk et al., 2020). In the case of forecasting COVID-19 in Germany and Poland, Bracher et al. (2021a) state that they did not find any “striking patterns” between modeling strategies in their analysis, but also acknowledge that this might be due to the relatively short study period they considered (10 weeks). We thus want to investigate whether, given a longer study period, patterns can be found.

We thus investigate performance at the level of the different modeling strategies as defined in section 2.1. To this end, we employ the method of pairwise comparisons as introduced in section 3.2.5 and apply them at the level of modeling strategies. Typically, the recommendation is to only compare models which are available for at least 50% of the time period under study, to avoid comparison of forecasts that don’t have any overlap (Bosse et al., 2022a). Since we are comparing forecasts at the level of model types rather than single models and there does not exist a pair of model types without overlap in the relevant time period, this advice does not necessarily apply here. One could still make the argument that one should not include models in the comparison that only contribute for a very small portion of the study period, so as not to have models that are perhaps not as representative of a modeling strategy[12] influence the results too much - however, we found that the results were not at all sensitive to the exact choice of availability threshold chosen for the models, so we decided not to exclude any models here.

Nevertheless, there remains the issue of the more obscure model types: as previously stated, the availability of agent-based and expert-judgment based models is low in our dataset. Specifically, the two agent-based models in the set only submitted forecasts for Poland, while the expert-judgment based models dropped out of the Hub during the late summer period of 2021. We report their results at the highest level of aggregation (averaged results across all locations and forecast dates), but will leave them out in the subsequent comparisons, for conciseness and as we think that their performance is not as representative of a general modeling strategy. For an in-depth evaluation and discussion of the relative performance of forecasts based on human judgment, we refer to Bosse et al. (2021).

The results of the pairwise comparison, where for each model type, weighted interval scores are stratified by horizon and target type, but averaged across the entire study period and all locations, can be found in Figure 8. These plots show that the only modeling strategy that tends to show increased performance

[12] One could think that models that only have e.g. 10% availability either dropped out due to not performing well and/or teams might have not been invested enough in the project to update and keep tuning their model, both of which might give low-quality predictions that as a result might not necessarily be representative of the respective model type.



Target: Cases

1 week ahead

	semi	mechanistic	baseline	statistical	expert judgment	agent−based
agent−based	0.57	0.62	0.58	0.6	1.12	1
expert judgment	0.88	0.74	0.78	0.87	1	0.9
statistical	1.05	1	0.91	1	1.15	1.66
baseline	1.16	1.09	1	1.1	1.28	1.72
mechanistic	1.06	1	0.91	1	1.35	1.61
semi	1	0.95	0.87	0.95	1.13	1.74

2 weeks ahead

	semi	mechanistic	baseline	statistical	expert judgment	agent−based
agent−based	0.49	0.56	0.65	0.6	0.77	1
expert judgment	0.9	0.75	0.92	0.86	1	1.3
statistical	0.88	0.99	0.95	1	1.17	1.67
baseline	0.93	1.04	1	1.05	1.08	1.54
mechanistic	0.89	1	0.96	1.01	1.33	1.8
semi	1	1.12	1.07	1.14	1.12	2.05

3 weeks ahead

	semi	mechanistic	baseline	statistical	expert judgment	agent−based
agent−based	0.4	0.48	0.73	0.58	0.42	1
expert judgment	0.81	0.62	1.03	0.83	1	2.35
statistical	0.71	0.99	1.03	1	1.21	1.72
baseline	0.69	0.96	1	0.97	0.97	1.37
mechanistic	0.72	1	1.04	1.01	1.6	2.09
semi	1	1.39	1.45	1.42	1.23	2.47

4 weeks ahead

	semi	mechanistic	baseline	statistical	expert judgment	agent−based
agent−based	0.31	0.41	0.77	0.53	0.33	1
expert judgment	0.73	0.54	1.13	0.84	1	3.06
statistical	0.56	0.95	1.13	1	1.2	1.89
baseline	0.5	0.84	1	0.89	0.89	1.29
mechanistic	0.59	1	1.19	1.05	1.84	2.46
semi	1	1.68	1.99	1.79	1.36	3.2

Target: Deaths

1 week ahead

	semi	mechanistic	baseline	statistical	expert judgment	agent−based
agent−based	0.91	1.04	0.88	0.86	1.52	1
expert judgment	0.84	0.82	0.78	0.98	1	0.66
statistical	0.82	0.84	0.88	1	1.02	1.16
baseline	0.93	0.96	1	1.14	1.28	1.13
mechanistic	0.98	1	1.04	1.19	1.22	0.96
semi	1	1.02	1.07	1.22	1.18	1.1

2 weeks ahead

	semi	mechanistic	baseline	statistical	expert judgment	agent−based
agent−based	0.8	1.1	0.79	0.78	1.46	1
expert judgment	0.98	0.97	0.69	0.89	1	0.69
statistical	0.83	0.94	0.87	1	1.13	1.28
baseline	0.95	1.08	1	1.15	1.44	1.27
mechanistic	0.88	1	0.92	1.06	1.03	0.91
semi	1	1.14	1.05	1.21	1.02	1.25

3 weeks ahead

	semi	mechanistic	baseline	statistical	expert judgment	agent−based
agent−based	0.66	0.94	0.7	0.66	0.94	1
expert judgment	0.93	0.96	0.65	0.82	1	1.06
statistical	0.75	0.95	0.92	1	1.23	1.51
baseline	0.81	1.03	1	1.09	1.54	1.43
mechanistic	0.79	1	0.97	1.06	1.04	1.06
semi	1	1.27	1.23	1.34	1.08	1.51

4 weeks ahead

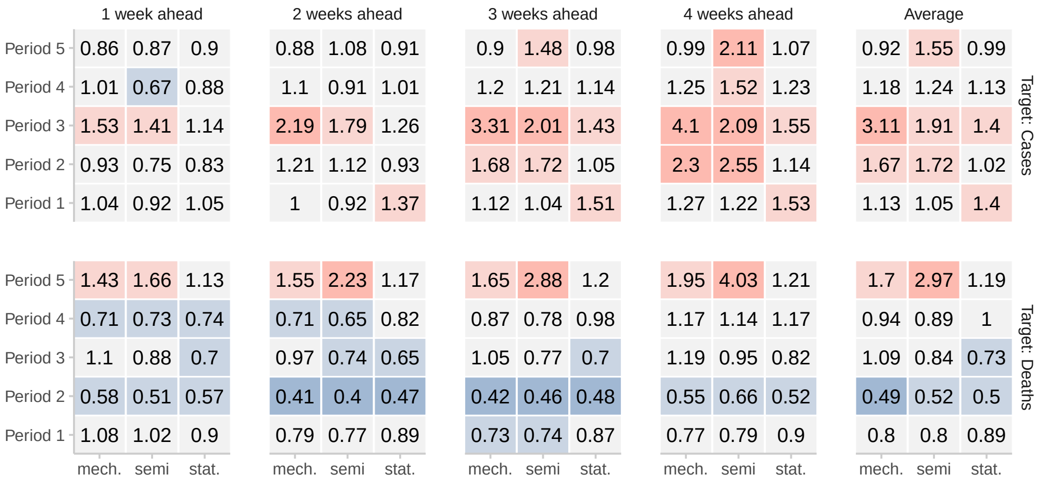
	semi	mechanistic	baseline	statistical	expert judgment	agent−based
agent−based	0.52	0.71	0.67	0.62	0.62	1
expert judgment	0.82	0.88	0.64	0.76	1	1.62
statistical	0.61	0.86	0.99	1	1.32	1.6
baseline	0.61	0.87	1	1.01	1.56	1.49
mechanistic	0.71	1	1.16	1.16	1.14	1.41
semi	1	1.42	1.64	1.65	1.22	1.93

Figure 8: Mean score (WIS) ratios resulting from pairwise comparisons of the modeling strategies in the European Forecast Hub, for forecasting incident cases and deaths during the time period March 2021 - January 2022. Comparisons are split up by the target series and forecast horizon. Scores are averaged across the five locations considered in this study. Values below one (indicated by blue tiles) mean that the modeling strategy in the given row on average performed better than the modeling strategy in the respective column, for the given forecast horizon and target series. Correspondingly, values above one (indicated by red tiles) mean that it performed worse. The plot was produced with the `scoringutils` package (Bosse et al., 2022b).

with respect to all other model types is the category of agent-based models. It appears that, compared to other model types, they perform particularly well for forecasting case numbers, even more so at longer horizons. For deaths, they perform similarly at lower horizons, but again show dominance over all other modeling strategies at longer horizons. As previously mentioned, within our data set, these models only issue forecasts for Poland (for both the cases and deaths series), so the comparisons are purely based on data from that location. Thus, it is unclear whether their superior performance would transfer to other locations or settings. In particular, the other groups of model types generally contain several models that issue forecasts for multiple locations, and one could argue that designing and tuning a model for a specific location lacks some of the potential complications that arise when aiming to establish a model with more universal application. We would thus argue that comparability is somewhat limited.

Apart from this, mean score ratios mostly stay close to one, with some slight variations. In particular, statistical models, as well as the approaches based on expert judgment, perform similarly to or even slightly better than semi-mechanistic or mechanistic models. At least in the aggregate, we thus cannot confirm that incorporating epidemiological principles gives an advantage in predictive performance. In fact, at horizons three and four weeks ahead, semi-mechanistic models are outperformed by both mechanistic and statistical models, more so for forecasting cases than for forecasting deaths.

However, as previously stated, aggregate scores are generally vulnerable to be dominated by a small number of single targets that resulted in higher scores. It could thus be that this result only emerges from certain locations or time periods with higher incidence levels. Conversely, it could also be the case that while there are no consistent trends in the aggregate for either of the two series, there might be systematic differences in scores at lower resolutions. Put differently, it is perhaps overeager to assume that one model type could systematically outperform another over the entire study period, which, after all, comprises different countries with varying periods of infection dynamics - however, patterns might



		1 week ahead			2 weeks ahead			3 weeks ahead			4 weeks ahead			Average		
		mech.	semi	stat.	mech.	semi	stat.	mech.	semi	stat.	mech.	semi	stat.	mech.	semi	stat.
Target: Cases	Period 5	0.86	0.87	0.9	0.88	1.08	0.91	0.9	1.48	0.98	0.99	2.11	1.07	0.92	1.55	0.99
	Period 4	1.01	0.67	0.88	1.1	0.91	1.01	1.2	1.21	1.14	1.25	1.52	1.23	1.18	1.24	1.13
	Period 3	1.53	1.41	1.14	2.19	1.79	1.26	3.31	2.01	1.43	4.1	2.09	1.55	3.11	1.91	1.4
	Period 2	0.93	0.75	0.83	1.21	1.12	0.93	1.68	1.72	1.05	2.3	2.55	1.14	1.67	1.72	1.02
	Period 1	1.04	0.92	1.05	1	0.92	1.37	1.12	1.04	1.51	1.27	1.22	1.53	1.13	1.05	1.4
Target: Deaths	Period 5	1.43	1.66	1.13	1.55	2.23	1.17	1.65	2.88	1.2	1.95	4.03	1.21	1.7	2.97	1.19
	Period 4	0.71	0.73	0.74	0.71	0.65	0.82	0.87	0.78	0.98	1.17	1.14	1.17	0.94	0.89	1
	Period 3	1.1	0.88	0.7	0.97	0.74	0.65	1.05	0.77	0.7	1.19	0.95	0.82	1.09	0.84	0.73
	Period 2	0.58	0.51	0.57	0.41	0.4	0.47	0.42	0.46	0.48	0.55	0.66	0.52	0.49	0.52	0.5
	Period 1	1.08	1.02	0.9	0.79	0.77	0.89	0.73	0.74	0.87	0.77	0.79	0.9	0.8	0.8	0.89

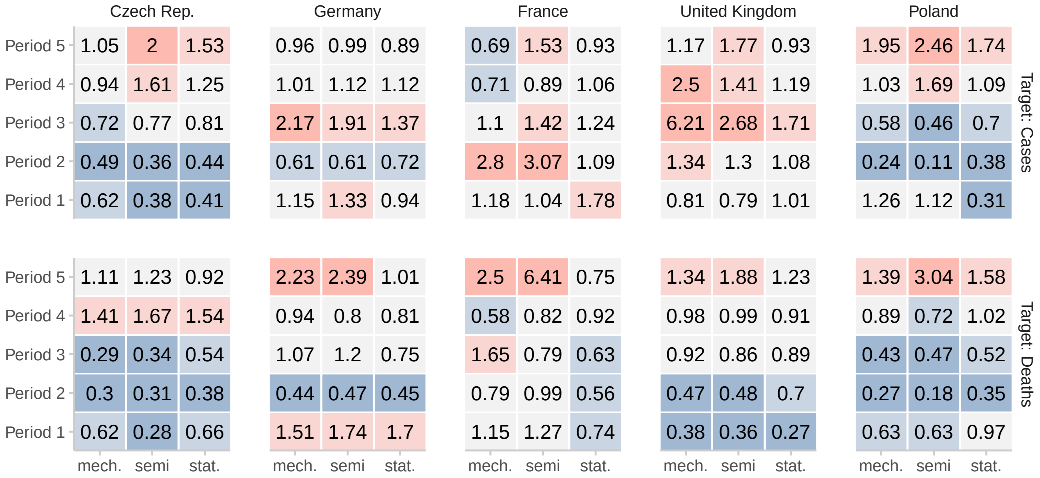
Figure 9: Relative WIS resulting from pairwise comparisons of the dominant modeling strategies (mechanistic, semi-mechanistic, statistical) against the baseline model in the European Forecast Hub, for forecasting incident cases and deaths during the time period March 2021 - January 2022. The time period under study is divided into two 10-week and three 9-week periods and comparisons are split up by the forecast horizon. Results are averaged across the five locations considered in this study. Values above one (indicated by red tiles) mean that the respective modeling strategy on average performed worse than the baseline model for the given period, forecast horizon and target series. Correspondingly, values below one (indicated by blue tiles) mean that it performed better than the baseline model. Lower relative WIS of one model strategy compared to another similarly correspond to better performance. The code to produce this plot was adapted from the `scoringutils` package (Bosse et al., 2022b).

exist at lower resolutions, that is, if comparisons are broken up by location or period. We investigate whether this is the case in the following.

We thus perform the same pairwise comparisons, once stratified by period and horizon in Figure 9, and once stratified by period and location in Figure 10. For conciseness, we refrain from directly showing all pairwise comparisons and only report performance against the baseline - as can easily be seen from equation 3.3, if one model (type) has a lower mean score ratio with respect to the baseline model than another model (type), it will also outperform it in terms of their direct mean score ratio.

First, consider Figure 9, which shows mean score ratios by period and forecast horizon (and averaged across locations). While in the aggregate we observed that the different model types did not substantially outperform the baseline, breaking comparisons up by period reveals that mean score ratios vary more, showing both examples of under- and overperformance with respect to the baseline model. In fact, it appears that there is actually more variation in mean score ratios with respect to the baseline between the different periods than between model types in a given period. That is, the model types often collectively outperform the baseline model (perhaps most prominently for forecasting deaths in period 2) or collectively underperform/stay close in performance to the baseline model. Put more concretely, it appears that some periods are generally easier/harder to forecast in comparison to the baseline, and that performance differences between the model types are still not that pronounced. The same applies for many periods for the pairwise comparisons by location in Figure 10.

There are however some exceptions: notably, in most periods we observe that as forecast horizon increases for the cases series, relative performance of the semi-mechanistic models deteriorates. While this is to a certain degree also true of the other model types (especially in period 3), semi-mechanistic models critically receive high relative scores in period 5. As this is the period where incidence of cases is highest



	Czech Rep.			Germany			France			United Kingdom			Poland		
	mech.	semi	stat.	mech.	semi	stat.	mech.	semi	stat.	mech.	semi	stat.	mech.	semi	stat.
Period 5	1.05	2	1.53	0.96	0.99	0.89	0.69	1.53	0.93	1.17	1.77	0.93	1.95	2.46	1.74
Period 4	0.94	1.61	1.25	1.01	1.12	1.12	0.71	0.89	1.06	2.5	1.41	1.19	1.03	1.69	1.09
Period 3	0.72	0.77	0.81	2.17	1.91	1.37	1.1	1.42	1.24	6.21	2.68	1.71	0.58	0.46	0.7
Period 2	0.49	0.36	0.44	0.61	0.61	0.72	2.8	3.07	1.09	1.34	1.3	1.08	0.24	0.11	0.38
Period 1	0.62	0.38	0.41	1.15	1.33	0.94	1.18	1.04	1.78	0.81	0.79	1.01	1.26	1.12	0.31
Period 5	1.11	1.23	0.92	2.23	2.39	1.01	2.5	6.41	0.75	1.34	1.88	1.23	1.39	3.04	1.58
Period 4	1.41	1.67	1.54	0.94	0.8	0.81	0.58	0.82	0.92	0.98	0.99	0.91	0.89	0.72	1.02
Period 3	0.29	0.34	0.54	1.07	1.2	0.75	1.65	0.79	0.63	0.92	0.86	0.89	0.43	0.47	0.52
Period 2	0.3	0.31	0.38	0.44	0.47	0.45	0.79	0.99	0.56	0.47	0.48	0.7	0.27	0.18	0.35
Period 1	0.62	0.28	0.66	1.51	1.74	1.7	1.15	1.27	0.74	0.38	0.36	0.27	0.63	0.63	0.97

Figure 10: Relative WIS resulting from pairwise comparisons of the dominant modeling strategies (mechanistic, semi-mechanistic, statistical) against the baseline model in the European Forecast Hub, for forecasting incident cases and deaths during the time period March 2021 - January 2022. The time period under study is divided into two 10-week and three 9-week periods and comparisons are split up by the locations considered in this study. Results are averaged across forecast horizons. Values above one (indicated by red tiles) mean that the respective modeling strategy on average performed worse than the baseline model for the given period, location and target series. Correspondingly, values below one (indicated by blue tiles) mean that it performed better than the baseline model. Lower relative WIS of one model strategy compared to another similarly correspond to better performance. The code to produce this plot was adapted from the `scoringutils` package (Bosse et al., 2022b).

for all locations, this is likely the root cause of these models comparing unfavorably for longer horizons in the previously discussed aggregate (Figure 8). For forecasting deaths, we only see this trend during period 5 - since these models however perform equally well to the other groups during period 1 (where incidence levels for deaths were also high, albeit falling), aggregate scores are not quite as affected as for forecasting cases.

Generally, the rankings that are induced by the mean score ratios for the model types mostly don't diverge between horizons. However, especially for forecasting cases, differences in performance often become more pronounced, suggesting that differences in performance get exacerbated with rising forecast horizon as forecasts diverge more from the target.

At the level of individual locations (Figure 10), we again observe that semi-mechanistic models perform worse than other model types in period 5 for forecasting both cases and deaths, although differences are more or less pronounced for the different locations and series - for instance, performance is very similar in Germany for forecasting cases, whereas these models receive an especially high relative score in France for forecasting deaths.

Furthermore, while examples of performance deterioration relative to the baseline are evident for all groups, these overall seem to be less pronounced for statistical models - the worst mean score ratio received by the group of statistical models is 1.78 when stratifying by location, while the other two modeling strategies receive substantially higher mean score ratios relative to the baseline at times.

Apart from this, we do sometimes see differences in relative skill between the model types, although these often change by period and no striking pattern seems to emerge - although we did attempt to, it is difficult to determine the cause of these differences, especially as the individual groups of models are sometimes quite small for some of the locations and individual models are often not available for the

entire (subset of) time period.
However, while these differences might not be easily explainable, it is a fact that they nevertheless exist at certain times and it thus might be beneficial for ensemble performance (in terms of the WIS) if an ensemble more heavily relied on some model types during certain periods. For example, if we consider the cases series for France, it might have been beneficial to, for instance, rely more on statistical models in period 2, and on the forecasts made by mechanistic models in all other periods. Conversely, for forecasting deaths in this location, statistical models seemed to issue the better predictions in all but period 4. We investigate the idea of ensemble composition based on modeling strategies more in section 6.2.

So far, we have discussed differences in performance only as it relates to the mean score ratios induced by the WIS. However, as also mentioned in our introduction of the methodology, there are other measures that can be used to assess modeling performance and that might give rise to different rankings.
As explained in subsection 3.2.4, the weighted interval score can straightforwardly be decomposed into penalties accrued from overprediction, underprediction and dispersion. A question that thus naturally arises is whether the score differences might stem from model types performing particularly well or not with respect to a certain component. To this end, we show the decomposed average WIS obtained by each model type, as well as the absolute error and measures of central interval coverage and bias, in Figure 11. Note that for the WIS and the absolute error, we report raw average scores rather than scores obtained relative to the baseline model.[13]
For the levels of overall WIS in panels (I) and (II), at lower horizons scores in general seem to be varying more between the different forecast horizons than they do between model types. At larger horizons, we once more observe that semi-mechanistic models receive relatively high aggregate scores, for both target series.
For forecasting cases, panel (I) shows that this model group receives substantially higher scores from the overprediction component of the WIS at larger horizons, while absolute error (panel (VI)) is also very high. However, according to panel (V), they also show some downward bias, which increases with horizon. Recall that the bias measure, contrary to the WIS, does not scale with the absolute level of the target series and reflects an overall tendency of *relative* under- or overprediction. This thus suggests that as a group, these models actually tend to slightly underpredict case numbers, but then accrue especially high absolute overprediction scores, likely during times of high incidence. Combined with what we discussed previously, that is, the high relative WIS this model group had during period 5, this suggests that these models overshot the target during this period, while generally underpredicting during other times. This trend of overshooting during peaks is a behavior that was also described by Bracher et al. (2021a) for models based on growth rate approaches. A possible reason for this could be that these models are more vulnerable to overprediction at longer horizons, as (slight) overprediction of the growth rate or the reproduction number, through the multiplicative nature of the underlying epidemiological process, can lead to large absolute overpredictions of the target.
In other cases, the tendency for over- or underprediction as induced by the WIS decomposition and the bias metric are more in agreement. For forecasting deaths, both mechanistic and semi-mechanistic models receive higher scores from the overprediction component than the underprediction component, while also showing very slight upward bias. Conversely, statistical models overall receive higher scores for the underprediction component, while also showing negative bias overall, especially for forecasting cases.
Furthermore, panels (III) and (IV) show that semi-mechanistic models tend to be better calibrated,

[13] We again account for changing group size over time by first averaging group results by forecast dates.

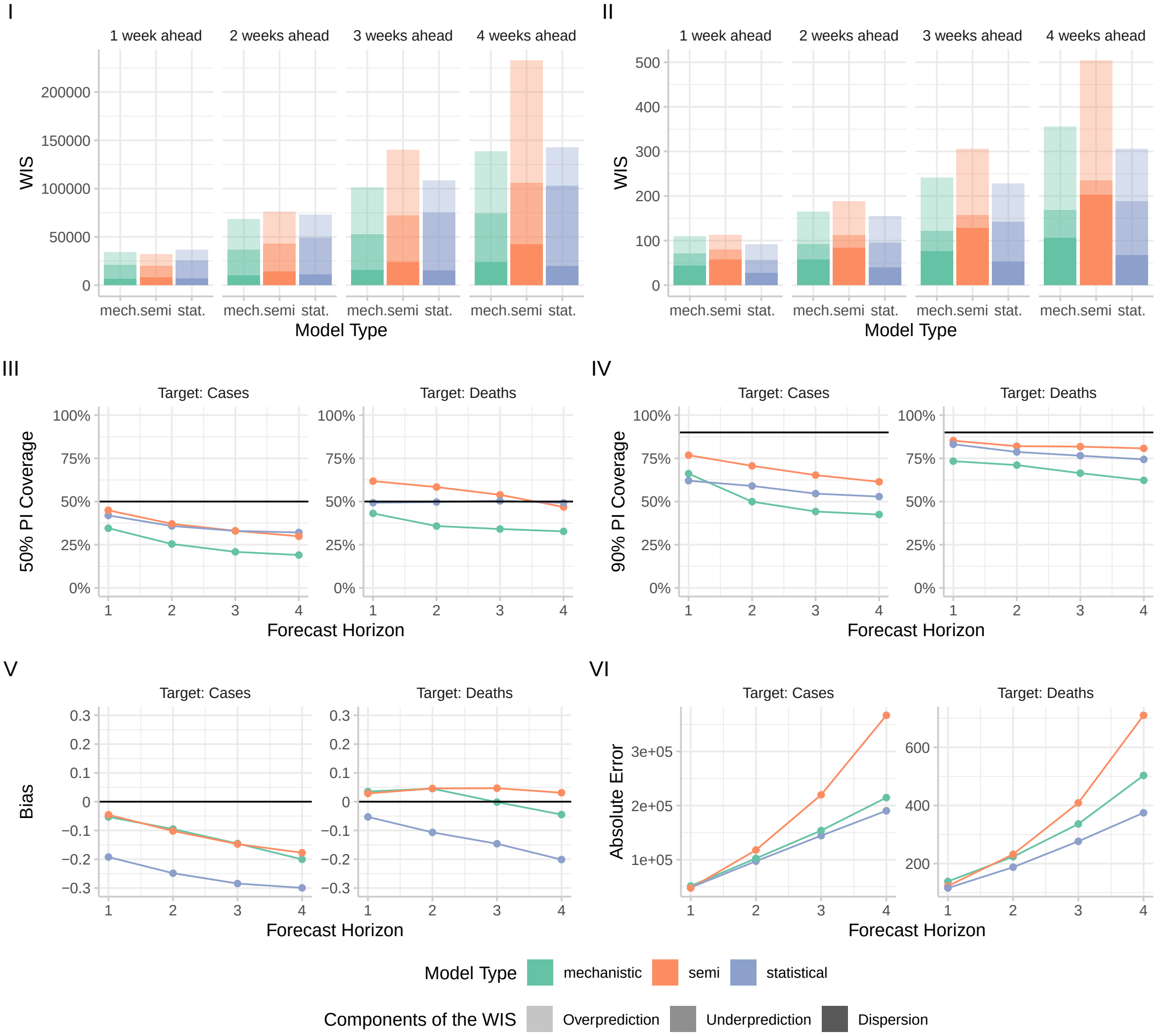


Figure 11: Performance with respect to several evaluation metrics of the three dominant modeling strategies (mechanistic, semi-mechanistic, statistical) in the European Forecast Hub, for two targets (case and death forecasts) and forecast horizons one to four weeks into the future. Results are averaged across five locations, for forecasts made during the time period March 2021 - January 2022. Respectively, the depicted scoring rules / evaluation metrics are: (I), (II) Decomposition of the weighted interval score into overprediction, underprediction, dispersion. (III) Empirical coverage of the 50% prediction interval. (IV) Empirical coverage of the 90% prediction interval. (V) Forecast bias, which ranges between -1 and 1. (VI) absolute error of the median forecasts. Wherever applicable, the desired target value of the score is shown as a black solid line - for all other scoring rules, a lower value of the score corresponds to better performance. Plot heavily inspired by Bosse et al. (2021).

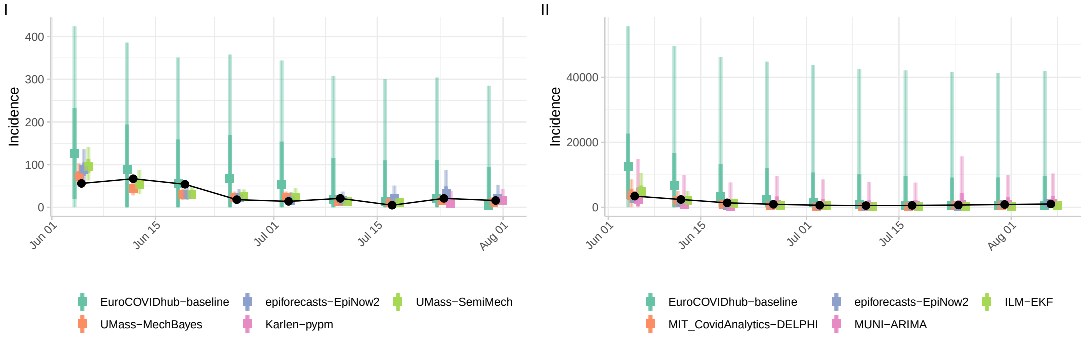


Figure 12: Two week-ahead forecasts for incident deaths in the Czech Republic (I) and incident cases in Poland (II), for the baseline model as well as four (randomly chosen) component forecasts, during a subset of the period under study that was marked by decline and overall low levels for the respective series. The black line and points show the realized true value of the respective series, while colored lines and points show the forecasts's median prediction, as well as the central 50% and 90% prediction intervals.

that is, show closer to nominal coverage rates than other model types - this is especially true at the 90% level for both targets, and at the 50% level for case forecasts, while these models tend to be more under-confident for death forecasts at the 50% level, where the group of statistical models is almost perfectly calibrated. This shows a common trade-off: as shown by the decomposition of the WIS in panels (I) and (II), semi-mechanistic models lack sharpness for forecasting both target series, but this might allow them better coverage, while other model groups might issue sharper forecasts at the expense of not containing as many of the observations as they are meant to. As stated in Sherratt et al. (2022b), the ensemble in the European Hub suffers from increased overconfidence for forecasting cases with rising forecast horizon on top of general sub-nominal coverage levels, which suggests that these models still might enter favorably into the ensemble by helping to counteract that overconfidence - Bracher et al. (2021a) made a similar argument for a single component forecast with large dispersion.
Generally, we however see a trend of overconfidence for all model types for both targets and furthermore we see a bit of a downward trend in the coverage ratio with rising horizon, especially for case forecasts. This is in line with the general results for component models in the European Hub, as described by Sherratt et al. (2022b).

We shortly discuss two more general observations that are not directly related to the issue of comparing modeling strategies. Recall what we previously stated about some periods being easier to forecast in comparison to the baseline: this is especially obvious in period 2 for forecasting deaths, where the series is mostly marked by decline and low levels overall (see Figure 5) and we observe that all modeling strategies markedly outperform the baseline model at virtually all horizons and locations. The consistency of this result across the locations (as well as for some locations for the cases series) led us to investigate the decomposition of the baseline model's WIS during this period. We found that its dispersion component alone sat between 0.96 (at horizon 1) and 1.46 (at horizon 2) of the *overall* average WIS of all other model types for forecasting deaths, suggesting that the baseline model issued severely under-confident forecasts during this time. Two examples of this behavior are shown in Figure 12, for the deaths series in the Czech Republic in panel (I) and the cases series in Poland in panel (II), where it is evident that the baseline model's predictive distributions are far wider than those of other (randomly chosen) component models. Since the baseline model's uncertainty bands are based on past differences, this suggests that these might not update quick enough in a situation where incidence is low after a period of higher levels, where all other modeling strategies are generally more capable of issuing more confident forecasts. This marks the point that the choice of baseline is not entirely trivial and that both individual models as well as the more general modeling strategies considered here might appear more capable if only compared to the baseline during this period, rather than to other models.

In relation to aggregate scores potentially being dominated by individual locations, consider the results from case forecasts in period 2 in Figure 10 and the average in Figure 9: it becomes evident that relative performance to the baseline, as well as the ranking between the modeling strategies, is critically influenced by France and the United Kingdom, which saw higher levels of incidence during this time. Whether or not this is a desirable feature of the evaluation method is an ongoing debate and can depend on the forecasters' as well as decision makers' preferences. For instance, Bracher et al. (2021a) argue that scoring forecasts in this manner is meaningful, as a fixed relative deviation from the observed quantity can be regarded as more problematic at high incidence levels rather than at very low incidence levels, which would not be accounted for if only considering the relative deviation. However, we would argue that especially in a situation such as here, where we are considering somewhat heterogeneous locations, this feature of scores can be misleading, as it presents results in the aggregate that can stem from diverging results at lower resolutions.

Overall, we conclude from this section that we can't reliably establish a consistent ranking between the three modeling strategies considered here, as all showed periods where they both over- or underperformed with respect to the other strategies (as well as the baseline model), and moreover no clear pattern emerged across the dimensions of the locations, target series or different periods. Upon further investigations, we additionally found that variability within these categories was often sizable. Thus, while the modeling strategies are distinct in their approaches and employed toolkits to base short-term forecasts on, it does not seem that one approach is fundamentally better suited than another in our application. On the flip side, this also means that explicit assumptions about the epidemiological process as well as modeling of transmission dynamics don't strictly seem to be necessary for short-term forecasting of COVID-19. In fact, there are many examples of statistical models performing slightly better than the other two groups, for example for forecasting deaths during the last period in our study.

One somewhat consistent result across the locations considered here (and that was identified previously in Bracher et al. (2021a)) is the relatively poor performance of the group of semi-mechanistic models for forecasting cases during the period that saw the most growth, as well as for longer horizons in general. At least for case forecasts, we however must also again note here that while this behavior is of course in principle undesirable, forecast performance has generally been found to be unreliable after one or two weeks into the future anyway, so it might not do well to dwell too much on differences occurring at longer horizons.

An initial goal we had in mind for this analysis was the potential to find patterns across modeling strategies that could be leveraged for ensemble composition - while we did not succeed in finding any patterns that are neatly aligned with certain locations or periods of the study, we did observe that substantial differences in performance still occurred at times. In section 6.2, we thus want to investigate whether automatic weighting schemes that are implemented at the level of modeling strategies can improve the performance of an ensemble.

5.2 Adding a new model to an ensemble

In this section, we take a more experimental approach, to analyze how ensemble performance can be impacted by the addition of single component forecasts.

While it is a universal result that ensemble models generally exhibit more robust performance than single models, little is known about how individual models can affect an ensemble's performance. This can however be a relevant question: consider the situation where, given an already established ensemble, one is proposed a new model and is consequently confronted with the decision of whether to add it to the base ensemble. The current practice within the Hubs is to include all models that pass a series of data format validity checks - although Ray et al. (2021) state that forecasters that appeared to be outliers used to be manually removed before the U.S. Hub switched from the mean ensemble to using a median ensemble. However, we believe that this decision of inclusion could be more crucial for smaller base ensembles as they exist in the European Hub, even when using the more robust median ensemble: at smaller ensemble sizes, the aggregation function of the ensemble is presumably less robust to the performance of individual models.

This experiment is somewhat inspired by the analysis in Bosse et al. (2021): they investigated the effect of adding three specific models to the Hub ensemble in Poland and Germany, and found that adding or removing a model induced performance changes that were "of a similar order of magnitude" between the mean and median ensemble, prompting us to consider the general effects of adding a model to an existing ensemble. There are several different lines of inquiry that thus arise with respect to this. First of all, it is interesting to consider whether it is more "safe" to add a model during certain periods than others, especially for periods with higher incidence levels. Furthermore, effects might be different based on the type of model added to the ensemble, with regard to models being better or worse performers or more or less distant to the base ensemble. Lastly, all of these effects might be different for the mean and median ensemble.

Concretely, we decided that the most systematic path to investigate these questions was via first building alternative ensembles of constant size - that is, at each of the available forecast dates, we recombined the set of available models into all possible sets of size k, which we subsequently aggregated into equally weighted mean and equally weighted median ensembles. We believe these "alternative ensembles" and their corresponding performance to be valid counterfactuals, as we generally believe that participation of one model at a given forecast date is independent of that of another. We thus obtain a large sample of alternative ensembles, thus allowing us to investigate other effects while removing the confounding factor of ensemble size.

Based on these data, we then designed the following experiment: we took a random sample of size 100 from all possible ensembles with k member models at each forecast date, location and for each of the target series. For k, we decided to investigate a relatively small base ensemble size of $k = 4$ as well as a more moderate size of $k = 8$. Due to the differences in model availability, the latter setting was only viable for Germany and Poland.

For each of these base ensembles, we proposed four of the remaining component forecasts at the given forecast date as new members for the ensemble. We deliberately kept the number of proposed models constant as we did not want to skew results by fluctuations in model availability over time.[14] From this set of proposed models, we added both the best and worst recent performer (as judged by the relative WIS obtained on the as of the particular forecast date resolved forecasts of the past four weeks), as well as the model with the largest as well as smallest distance (as judged by the Cramer distance in

[14] Although we still had to exclude a few forecast dates for some locations where model availability was too low.

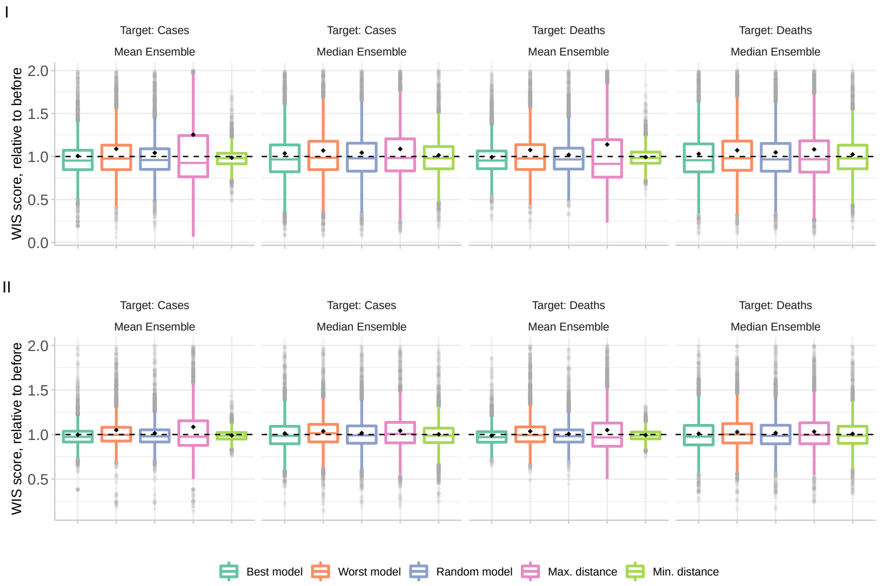


Figure 13: Distribution of scaled WIS after adding different types of models from a proposed set of four models to base ensembles with four (I) and eight (II) member models. Models were added based on the characteristics: best/worst performer, minimum/maximum distance to the base ensemble and a reference (randomly chosen) model. The WIS after adding the respective model was divided by the WIS of the base ensemble beforehand, as a measure of change in performance resulting from adding the respective model. Values below 1 thus correspond to an improvement after the addition, while values above 1 correspond to worsened performance. Outlying values > 2 are excluded from the plots, and the mean values of the scaled WIS (indicated in black) are included to account for this.

section 3.2.6) to the current ensemble prediction. Finally, we also added a random model from the set, as a reference to compare the resulting effects against. All of the resulting ensembles were subsequently scored via the WIS and divided by the base ensemble's score. We thereby get a direct measure of the change in performance as a consequence of adding the respective model.
We did this separately for the two ensemble types - note thus that the model added based on distance to the base ensemble was not necessarily the same for the mean and median ensemble, although we found that it was in most cases (e.g. for $k = 4$, there was an agreement of 76.5% for the minimum and 88.3% for the maximum distance model and for $k = 8$, an agreement of 72.2% for the minimum and 83.5% for the maximum distance model). We regard this as more of a sanity check of our implementation, as we would expect the mean and median ensemble to generally be closer to one another than to any component forecast and hence be closest to and (in particular) most distant to the same model. Conversely, the selection of the best or worst recent performer only depends on the set of proposed models, hence the model added based on the performance measure was always the same, regardless of ensemble type.

Figure 13 shows the distribution of relative scores, separately for the two target series, the different types of models added and the type of base ensemble. For both sizes of the base ensemble, the median relative scores for all added models tend to be closely below one, meaning that it was beneficial (albeit only slightly) to add another model in most cases, irrespective of that model's nature, the target series or the type of base ensemble. Importantly, this also means that even a model that has shown poor performance in recent weeks will benefit an ensemble's score in most cases. However, median and average relative scores for the best model tend to be a bit lower than those of the worst and the random model, suggesting that there might be some benefit in selecting for better recent performers when adding a model to an

ensemble.
Considering overall differences in variation between models for the mean ensemble, these are somewhat more pronounced: especially for $k = 4$, the maximum distance model is associated with a larger spread in scores, compared to the minimum distance model, which overall has comparatively little effect on performance. Furthermore, the median and average scores diverge substantially for the maximum distance model, suggesting that adding a model with large distance to a mean ensemble is simultaneously quite beneficial in most cases and can have large detrimental effects in others. Conversely, spreads in scores are overall more similar for the median ensemble, with only slightly larger variation in relative scores for the maximum distance model.
Differences in the median values or the spread of relative scores between the two target series are overall not apparent. Pertaining to differences with respect to the size of the base ensemble, spread in relative scores is consistently reduced for the larger setting, supporting our previous point that the performance of ensembles that already contain a larger number of component models becomes more robust to newly added ones of any nature. This is also supported by the average relative scores generally moving closer to one.

While the median relative scores thus mostly stay close below one, mean relative scores are often substantially higher. One could argue that these mean effects are also of considerable interest, as they give a more accurate picture of how "safe" it is to add a model to an ensemble. With the relative scores generated here, single outliers are less of an issue, measurements are not systematically correlated and the data set is considerably larger than in the original data set with direct absolute scores. We thus felt it was viable to fit a generalized linear model (GLM) - as we observed the conditional distributions of the outcome to be positively skewed (Figure A1) and as relative scores are restricted to the positive real line, the natural choice was to model scores via the two parameter gamma distribution as it is implemented in the `gamlss` package (Rigby and Stasinopoulos, 2005). This particular implementation models the distribution via two parameters, μ and σ, with μ the mean of the response and σ^2 the dispersion[15] parameter. The distribution's standard deviation linearly scales with the σ-parameter (which is modeled via a log-link function), meaning that a negative sign of an estimate for the σ-parameter corresponds to a ceteris paribus decrease of the standard deviation of the response. While it is not straightforward to interpret the size of the resulting coefficients, we can analyze whether the effects are of similar orders of magnitude for the different models and ensemble types.
Hence, we fit a regression of the relative score on the type of model that was added to the ensemble (best/worst recent performer; minimum/maximum distance), with the randomly added model as a reference category, separately for the two ensemble types. We additionally included the forecast horizon as a control, as well as random effect terms for the interaction of location, target series and the periods as defined in section 4.2, to control for idiosyncrasies that might have arisen at certain location - time combinations of the respective target series.
The resulting coefficients, as well as their standard errors, for $k = 4$ are reported in Table 1.[16] Pertaining to the recent performance of the added model, it can be observed for both ensemble types that this does have an effect on the average relative performance: adding a model that has shown good relative performance in the near future leading up to the forecast date on average improves performance more than a random choice of model, while adding a model that has performed relatively poorly also leads on

[15] Not to be confused with the dispersion component of the WIS.

[16] We refrain from explicitly marking the significance levels of estimates, as all estimates are highly significant. Due to the large sample we created (n = 535980) through the recombination experiment, we would also argue that the concept of statistical significance is not that meaningful here.

	μ		σ	
	Median Ens.	Mean Ens.	Median Ens.	Mean Ens.
β_0	1.0332	1.0369	-1.3150	-1.4938
	0.0013	0.0009	0.0030	0.0030
Best Model	-0.0147	-0.0250	-0.0046	-0.1289
	0.0014	0.0010	0.0030	0.0030
Worst Model	0.0213	0.0333	0.0282	0.1580
	0.0014	0.0012	0.0030	0.0030
Max. Distance Model	0.0305	0.0892	0.1120	0.5044
	0.0015	0.0016	0.0030	0.0030
Min. Distance Model	-0.0224	-0.0240	-0.2183	-0.8036
	0.0012	0.0008	0.0030	0.0030
Horizon	0.0060	0.0028	0.0770	0.0714
	0.0004	0.0002	0.0008	0.0008

Table 1: Coefficients from a fitted GLM with gamma-distributed outcome. Standard errors are reported below coefficients. The outcome variable is the WIS after adding different types of models from a proposed set of four models to base ensembles, relative to the WIS of the base ensemble before the addition. Base ensembles consisted of four member models, and effects were assessed separately depending on the type of base ensemble (quantile-wise median or mean ensemble). Models were added based on the characteristics: best/worst performer, minimum/maximum distance to the base ensemble and a reference (randomly chosen) model. Dummy effects were added for the type of model, with the random model as base category. Forecast horizon was added as control, and random effects were included to control for location, target type and period. For this specification, values above 3 were excluded. Model was fit using the `gamlss` package (Rigby and Stasinopoulos, 2005) - the μ parameter corresponds to the mean of the response distribution, σ^2 to its dispersion. Importantly: $N = 268990$, size of standard errors are thus not that meaningful.

average to a worse performance of the ensemble. In regards to the differences between the two ensemble types, we note that the effects for the mean are larger - moreover, the effects on the dispersion (and thus the variance) of the relative scores are substantially larger for the mean ensemble. This is likely due to the fact that the mean ensemble is more vulnerable to outliers - it thus seems to be the fact that it is particularly "safe" to select an additional model based on recent performance for the mean ensemble, compared to adding a random model.

For both ensemble types, variance of the relative scores is reduced when adding a model that is particularly similar to the established ensemble, compared to adding a random model. This is likely also exacerbated by the relatively small ensemble size we are considering here - at only four component models in the ensemble, adding another model can in principle greatly move the ensemble's predictions (and thus impact its performance). Adding a model that is more in agreement with the established ensemble thus leads to less of a spread in relative scores, while adding one that "disagrees" with will have larger impact. In theory, this could also mean that there is more potential for improvement. However, especially for the mean ensemble, adding a model with larger distance is also associated with an expected increase in scores, compared to simply adding a random model. When interest thus lies in stability, this suggests that care should be taken when considering an additional model that is in disagreement with the current ensemble.

Furthermore, across the board, we observe that effect sizes for the different models are larger for the mean ensemble, for both the mean and the dispersion parameter. This again supports the statement that the mean ensemble is moved more easily by an additional model and is in particular more vulnerable to outlying forecasts.

For the horizon effect, we generally see that average relative scores increase with the forecast horizon,

	μ		σ	
	Median Ens.	Mean Ens.	Median Ens.	Mean Ens.
β_0	1.0410	1.0466	-1.2423	-1.3739
	0.0018	0.0013	0.0038	0.0038
Best Model	-0.0218	-0.0349	0.0014	-0.1305
	0.0019	0.0014	0.0038	0.0038
Worst Model	0.0303	0.0544	0.0308	0.1890
	0.0020	0.0017	0.0038	0.0038
Max. Distance Model	0.0424	0.1446	0.1233	0.5693
	0.0021	0.0024	0.0038	0.0038
Min. Distance Model	-0.0272	-0.0277	-0.2641	-0.9580
	0.0017	0.0012	0.0038	0.0039
Horizon	0.0069	0.0019	0.0805	0.0630
	0.0005	0.0003	0.0011	0.0011

Table 2: Coefficients from a fitted GLM with gamma-distributed outcome. Standard errors are reported below coefficients. The outcome variable is the WIS after adding different types of models from a proposed set of six models to base ensembles, relative to the WIS of the base ensemble before the addition. Base ensembles consisted of four member models, and effects were assessed separately depending on the type of base ensemble (quantile-wise median or mean ensemble). Models were added based on the characteristics: best/worst performer, minimum/maximum distance to the base ensemble and a reference (randomly chosen) model. Dummy effects were added for the type of model, with the random model as base category. Forecast horizon was added as control, and random effects were included to control for location, target type and period. Model was fit using the `gamlss` package (Rigby and Stasinopoulos, 2005) - the μ parameter corresponds to the mean of the response distribution, σ^2 to its dispersion. Importantly: $N = 141732$, size of standard errors are thus not that meaningful.

which is likely due to the larger disagreement that forecasts generally exhibit with rising horizon.

To assess the results' sensitivity to large outlying relative scores, we reran the regression with relative scores > 3 excluded (Table A1). Effects remain, although their sizes tend to shrink slightly (especially for the mean ensemble and maximum distance model).

Furthermore, we expected that effect sizes would be bigger if we had more models to choose from, as the choice of a model based on both performance and distance measures becomes more meaningful if the set of models it is chosen from is larger. To check this, we reran the analysis for only Poland and Germany and consequently were able to propose two more models at each forecast date. The results are reported in Table 2, where we see that most effects are larger in the absolute, for both the mean and the dispersion parameter. In particular, the effects for the maximum distance model are increased. This suggests that the further a proposed model is from the base ensemble, the greater the risk for large deterioration in performance, compared to adding a random reference model. Effects on the mean and dispersion of the relative scores also increase in the absolute for the performance based models, suggesting that the better (or worse) a forecast model has performed in recent times, the safer (or riskier) it is to add to an ensemble.

Lastly, we report the regression results for the larger base ensembles ($k = 8$) in Table 3. In congruence with the reduction of variation in the aggregate relative scores that we discussed earlier, effect sizes are reduced for these larger ensembles. Nevertheless, the addition of the worst and maximum distance model still impact the average relative score, particularly for the mean ensemble. This suggests that at such an ensemble size, which is not uncommon for some countries in our set as well as other countries in the European Hub, ensembles can still be somewhat vulnerable to outlying or relatively poor performing member models.

	μ		σ	
	Median Ens.	Mean Ens.	Median Ens.	Mean Ens.
β_0	1.0110	1.0145	-1.9595	-2.0467
	0.0010	0.0007	0.0042	0.0042
Best Model	-0.0056	-0.0167	-0.0085	-0.1993
	0.0011	0.0008	0.0042	0.0042
Worst Model	0.0132	0.0211	0.0050	0.1932
	0.0011	0.0009	0.0042	0.0042
Max. Distance Model	0.0153	0.0320	0.0668	0.4602
	0.0011	0.0011	0.0042	0.0042
Min. Distance Model	-0.0090	-0.0109	-0.1291	-0.8419
	0.0010	0.0006	0.0042	0.0042
Horizon	0.0027	0.0003	0.1242	0.0666
	0.0003	0.0002	0.0012	0.0012

Table 3: Coefficients from a fitted GLM with gamma-distributed outcome. Standard errors are reported below coefficients. The outcome variable is the WIS after adding different types of models from a proposed set of four models to base ensembles, relative to the WIS of the base ensemble before the addition. Base ensembles consisted of eight member models, and effects were assessed separately depending on the type of base ensemble (quantile-wise median or mean ensemble). Models were added based on the characteristics: best/worst performer, minimum/maximum distance to the base ensemble and a reference (randomly chosen) model. Dummy effects were added for the type of model, with the random model as base category. Forecast horizon was added as control, and random effects were included to control for location, target type and period. Model was fit using the `gamlss` package (Rigby and Stasinopoulos, 2005) - the μ parameter corresponds to the mean of the response distribution, σ^2 to its dispersion. Importantly: $N = 141840$, size of standard errors are thus not that meaningful.

Considering the question of whether it is in general safer/riskier to add models to an ensemble depending on the phase of the epidemic, we additionally show the fitted random effects from the model in Table 1 (Figure 14). The effects were contrast-coded, and thus represent deviations from the overall intercept. While variation between locations for a given period is somewhat pronounced for both target series, effects for the median ensemble (panel (I)) are overall more similar. For the mean ensemble (panel (II)) effects vary more and tend to be larger during period 5, particularly for the cases series. This could be due to the fact that models in general issue more varying predictions during these periods of rising incidence, especially for the cases series. This effect can also be observed for the median ensemble, albeit to a lesser degree, once more supporting the point that the median ensemble is more resistant to additions. For the setting with the larger base ensemble, variation in random effect sizes was substantially less pronounced, again supporting the point that larger ensembles become more robust with respect to individual models they include (Figure A2).

To conclude, while it emerged that adding another model will improve an ensemble's scores in most cases, our analysis also suggests that due to potentially large deteriorations in the relative score, care should be taken when adding a new model to an existing ensemble, especially when that model has shown poor performance in recent weeks or it is in disagreement with the current ensemble. When confronted with such a decision, it might be wise to manually check such a forecast model for plausibility - that is, whether e.g. its distance from the current ensemble could result from a (perhaps temporary) data or model misspecification error. It could also be beneficial to assess the performance of the current ensemble in recent weeks and appraise whether the proposed model could somehow "correct" undesirable characteristics of the ensemble - for instance, Bosse et al. (2021) identified a case where a poorly performing model nevertheless improved the overall median ensemble, as it was more "directionally correct" relative

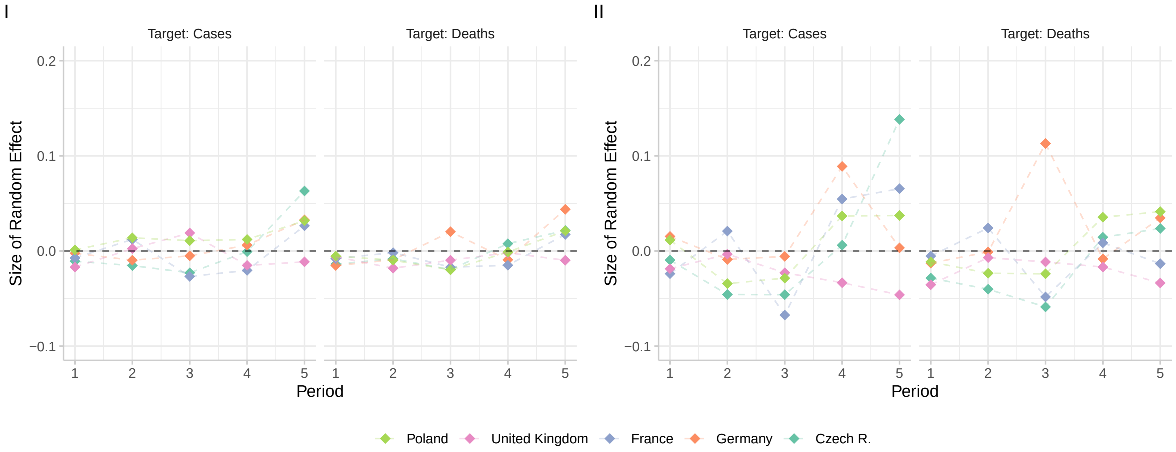


Figure 14: Fitted random effects for the mean parameter from a generalized regression with gamma distributed response. The response variable is the relative score (WIS) that results after adding a model to a base ensemble with four member models, using the quantile-wise median (I) or mean (II) as the aggregation function. The base ensembles were recombined from the set of available forecast models in the data at each forecast date, separately for the different target series and locations.

to it. Nevertheless, as the overall decrease in effect sizes for the larger base ensemble suggests, we would expect that at yet larger ensemble sizes (which we could not feasibly consider given our limited model base), ensembles are likely sufficiently robust to single models, such that manual screening of models might not be that beneficial.

Lastly, our results once more support the general result that the median is the more robust choice as an aggregation function for the ensemble, especially at the rather low to moderate number of member models we have considered here and during times of rising incidence, where individual model predictions are more varied.

Arguably, the analysis performed in this section was to a certain degree stylized. Moreover, it is not usual practice to consider model additions at single forecast dates. It would thus be interesting to consider the effect of an addition of a model to an ensemble not just at a single forecast date, but over a longer period of time. For our analysis, this was inhibited by the fact that (due to changing availability of component forecasts), ensembles consisting of a constant set of member models were often not available for longer periods of time. Nevertheless, the experiment allowed us to support some previously held intuitions, namely that both the recent performance of a model as well as its disagreement with the current ensemble can impact the performance of the ensemble it is a member of. However, while differences were consistently noticeable both between the types of added models and the type of base ensemble, it is difficult to judge how meaningful these differences in effect sizes are and how relevant they would be in practice. In the next section, we thus want to investigate whether "reverse-engineering" these results by changing the Hub ensemble composition such that it selects for better performers is a reasonable strategy.

6 Retrospective Ensemble Strategies

As stated before, the European Hub has so far mostly relied on the equally weighted median to build its official ensemble, and inverse score weighting based on past performance of all models has not shown a consistent improvement (Sherratt et al., 2022b). Nevertheless, the question remains whether the ensemble could in theory be improved by changing its composition, for instance by selecting certain forecast models and/or using a different weighting scheme.

In this section, we thus retrospectively analyze alternative methods to compose the ensemble, to investigate whether the way the ensemble is currently built for the European Hub can be improved. While this is a retrospective analysis, it is generally important that it, to the extent that it is possible, simulates the real-time situation, as it would otherwise be too easy to retroactively devise strategies that outperform the Hub ensemble.This is because we want to trial (and if successful, suggest) methods that could be deployed in real-time for a setting with smaller numbers of component forecasts and with at times incomplete records of historical performance.

We chose to compare the methods in this section to an ensemble that employed the unweighted median over the entire study period rather than to the actualized Hub ensemble. This is due to the fact that the aggregation method for the Hub ensemble was changed from mean to median in July of 2021. We noticed that some of our methods compared favorably to the actualized Hub ensemble during the early months of 2021. We felt that this was somehow misleading, so instead opted to use the median ensemble as a benchmark throughout the study period, after checking that our implementation of the ensemble gave similar predictions to the official Hub ensemble. Whenever we refer to the "benchmark" model during this section, we are thus referring to the equally weighted median ensemble that indiscriminately includes all models.

For reporting scores, we simply divide the respective method's average score by that of the benchmark. This is possible due to the fact that there are no missing forecasts for the ensembles we aim to compare, although we generally exclude the first few weeks, which are needed to estimate weights and/or select models.

6.1 Selection ensemble

In this section, we investigate whether changing the model base such that only the recent best performers are included in the ensemble, potentially in conjunction with additionally weighting component forecasts based on past performance, can lead to an improvement over the equally weighted median ensemble that indiscriminately includes all models. We have several reasons why we believe this approach could prove fruitful.

As stated, Sherratt et al. (2022a) showed that for the European Hub, employing a weighting scheme where all models are weighted based on their past performance did not provide a benefit over the equally weighted median, neither when subsequently using a weighted mean or weighted median. The question thus remains whether either pure model selection or combining weighting with prior selection of recent good performers could provide a benefit. In the case where it is possible to identify models that truly have an edge over others (or conversely: identify models that are consistently poor performers), it might be beneficial to compose a model set based on this information, whether or not models are subsequently weighted. In particular, estimating weights in a reduced ensemble could remove some noise from the estimation.

Furthermore, our results in section 5.2 showed that adding a relatively good performer on average has a better effect on the ensemble, compared with the addition of any model. We thus want to study whether this result can, somehow, be “reverse-engineered” by purposefully tuning the set of models such that poor performing models are left out and only the relatively good performers remain. Note that here the procedure is different to what we did in section 5.2: we are actively removing models from the set, while before we investigated the effect of adding a model to the ensemble. While these are to a certain degree just two sides of the same coin, in practice we would be more realistically faced with the decision of removing a model from the existing model base rather than adding an extra model to it.

Most importantly, the approach of selecting for good performers has some precedent in the U.S. Hub: Ray et al. (2022) find that a combination method that simultaneously reduces the model set to the ten recent best performers and subsequently weights these by their past performance (in terms of the relative WIS) performs consistently better than an equally weighted median for forecasting deaths, while it does not improve predictive performance for forecasting cases. As a consequence, the relative WIS weighted median ensemble superseded the equally weighted median ensemble as the U.S. Hub’s official method for forecasting deaths in November 2021.

In fact, they state that they also applied this method to the European Hub, where the method however showed worse performance relative to the equally weighted median including all component forecasts, for both deaths and cases. This could however be due to the substantial differences in model availability between the two Hubs - as previously stated, most models participating in the U.S. effort submit forecast for all locations (that is, states), while this is only the case for less than half the models in the European Hub. Since their approach only selects the best ten models measured by relative WIS across the entire European set, we surmise that it could fail due to two reasons.

First of all, the approach in effect substantially reduces the model set for most locations, as it will likely select some models which only submit forecasts for a small number of locations or even just a single location - for instance, recall our finding in section 5.1 that found that Poland’s agent-based models had substantially lower relative WIS than other model types, at least for forecasting cases, making it likely that these models would be selected. Hence, this way of selecting models presumably already leads to worse performance due to the sheer effect of greatly reducing the number of models in the ensemble at some locations. Ray et al. (2022) also mention that this could be an issue.

Second of all, we often noticed during our investigations of the data that a model that is a good performer for one location at a given time isn’t necessarily one for all other locations. Moreover, when using the WIS to judge component forecasts, we could have the case that locations with nominally high levels of incidence could dictate the model set for all locations. We thus believe that it’s better to allow for selection of models and weight estimation separately by location, even though this could also increase the variance in weight estimation.

Hence, we investigate whether an approach of treating each location’s model set separately can improve performance, as it ensures that each location is given a full set of best k performers unique to that location. Nevertheless, we are still dealing with the aforementioned issue that the model base is not large to begin with. So even when attempting to draw from the available resources as much as possible, reducing ensemble size further could counteract any potential benefits from choosing better performers, as the ensemble practically becomes more vulnerable to failure or erratic behavior of single component forecasts if the number of models is smaller.

Generally, due to the fact that Ray et al. (2022) extensively investigate these trained ensemble methods, we will often mention and discuss the results they obtained (for both the European and the U.S. Hub), to judge to what degree the results as well as the explanations they give for the potential drawbacks of the method might transfer to the European Hub.

We thus proceeded as follows. At each forecast date, via a rolling window of the respective past four weeks, we select the number of k models with lowest relative WIS within that window. Choosing models based on the relative WIS rather than average scores addresses the issue of potentially choosing models that only receive a low average score due to having missed "difficult" targets. Relative WIS is computed separately by location and target series, to mitigate the aforementioned issue of differences in the model set and/or performance by location.

If one of the selected models did not participate at the current date, the next best model takes its place - put differently, we only select from the set of models that are actually available at the current date, to ensure that ensemble size is always constant. Forecasts that have not resolved yet by the respective forecast date are excluded from scoring, since the resolution of, for instance, the most recent 3-week forecast would not be in the information set of the ensemble forecaster in a real-time setting. The set of scored targets thus includes one 4-week ahead forecast, two 3-week ahead forecasts, three 2-week ahead forecasts and four 1-week ahead forecasts.

As stated in section 3.3, the window size is of course a tuning parameter that can affect the results. Given that our analysis is entirely retrospective, we refrained from excessively tuning this parameter to the point where it would give the most favorable in-sample results. We therefore chose a similar value to the one used by Bracher et al. (2021b) in a related analysis - they used a window size of three weeks to estimate weights, which we slightly increased to four weeks as we wanted to include at least one forecast from all horizons. As stated, this parameter can be seen as a dial between the bias and variance of estimating model weights. We believe that four weeks strikes a good balance between the two.

The window size of four weeks also means that the first four weeks of our dataset serve as input for both the selection and weight estimation. We thus obtain a total of 43 forecast dates that can be evaluated, for all locations and target series.

As an aggregation method for the thereby induced sets of best recent performers, we considered the equally weighted median, the equally weighted mean, as well as a weighted approach for both. For the weighted methods, we employed inverse score weighting as detailed in section 3.3. In the cases where component forecasts had missed one target within the current window used to estimate the weights, we proceeded as in Bracher et al. (2021b) and imputed that past score with the worst score achieved by any model for that particular target.

We varied the number of component forecasts included, namely, we tried $k = 3, 5, 8, 10$. For a given k, we excluded a location and target type combination if the size of its model base was smaller than $(k+2)$ for at least half of the study period, as as we believed that in these cases the analysis would conceptually be too similar to an ensemble of all models (that is, there is no meaningful choice of "best" performers if almost all models are included over most of the study period) and could therefore portray the method more favorably than it is in actuality.

Before we turn to the results in performance, we want to shortly pull attention to Table 4, which, for $k = 5$, displays each location's top three chosen models using the selection method, separately for the two target series. We do observe that there are some models (most prominently `ILM-EKF` for Cases and `UMass-MechBayes` for Deaths) that are chosen across most or all locations. However, we also both see models (such as `epiforecasts-EpiNow2` or `MUNI-ARIMA`) that in principle submit forecasts for all locations, but are more often chosen at some locations than others, as well as some models, such as

	Location	Most chosen model (1st)	Most chosen model (2nd)	Most chosen model (3rd)
Cases	Czech R.	ILM-EKF (35)	IEM_Health-CovidP. (31)	MUNI_DMS-SEIAR (28)
	Poland	MOCOS-agent1 (29)	epif.-EpiNow2 (23)	ILM-EKF (21)
	Germany	itwm-dSEIR (25)	ILM-EKF (20)	ITWW-county_repro (16)
	France	ILM-EKF (34)	USC-SIkJalpha (32)	IEM_Health-CovidP. (31)
	U.K.	IEM_Health-CovidP. (34)	MUNI-ARIMA (27)	ILM-EKF (23)
Deaths	Czech R.	UMass-MechBayes (36)	ILM-EKF (28)	epif.-EpiNow2 (24)
	Poland	MOCOS-agent1 (37)	UMass-MechBayes (29)	ILM-EKF (27)
	Germany	HZI-AgeExtendedSEIR (28)	RobertWalraven-ESG (18) USC-SIkJalpha (18)	
	France	RobertWalraven-ESG (32) UMass-MechBayes (32)	IEM_Health-CovidP. (31) ILM-EKF (31)	
	U.K.	IEM_Health-CovidP. (26) USC-SIkJalpha (26)	LANL-GrowthRate (21) RobertWalraven-ESG (21) UMass-MechBayes (21)	

Table 4: Models that were most often chosen by the selection method, which bases ensemble composition based on recent model performance. For the results shown in this table, the best five component forecasts in the past four weeks (judged by the relative WIS) were included in the ensemble. Numbers in parentheses behind the model names indicate the number of times each model was chosen by the selection method. The third place is excluded wherever ties occurred in higher places. Names of IEM_Health-CovidProject and epiforecasts-EpiNow2 were shortened for formatting reasons.

ITWW-county_repro or MOCOS-agent1 that forecast for only one or two locations. Recall that we argued that the U.S. approach could be fundamentally flawed for an application in the European Hub as it does not account for the fact that models could have considerable variations in performance across locations. While it does seem to prove true that such performance differences exist and a more flexible method will accordingly choose different models, as we will shortly see, accounting for this is not enough to gain competitive performance relative to the benchmark in the European Hub.

We now turn to an actual discussion of the results. Table 5 shows the ratio of the average WIS obtained by the respective method and that of the benchmark, for $k = 5$ and $k = 10$. We furthermore show the variation of relative scores obtained at each forecast date, separately by location and the two target series, in the boxplots in Figure 15, for all k considered. For both, we additionally show the relative scores that result from averaging scores across all locations.

For the "average location", it is evident that for all numbers of component forecasters included in the best performers set, scores are consistently either similar to or worse than those of the benchmark, for all aggregation methods considered. Choosing a subset of best performers thus does not seem to be a viable alternative to the benchmark across locations, whether or not the models are subsequently weighted. Between locations, we however observe some variations and thereby also some examples of improved performance. We now discuss some interesting trends and results that emerge from the implementation, and additionally attempt to establish reasons/heuristics for why the approach tends to fail in comparison to the benchmark.

As a general trend, all approaches tend to work better the more models are included: Figure 15 shows that variability in performance mostly decreases and the median and average relative scores move closer to one with increasing k. This suggests that the positive effect of including more models in an ensemble outweighs the potential effect of selecting for better performers - the big exception here is the cases series

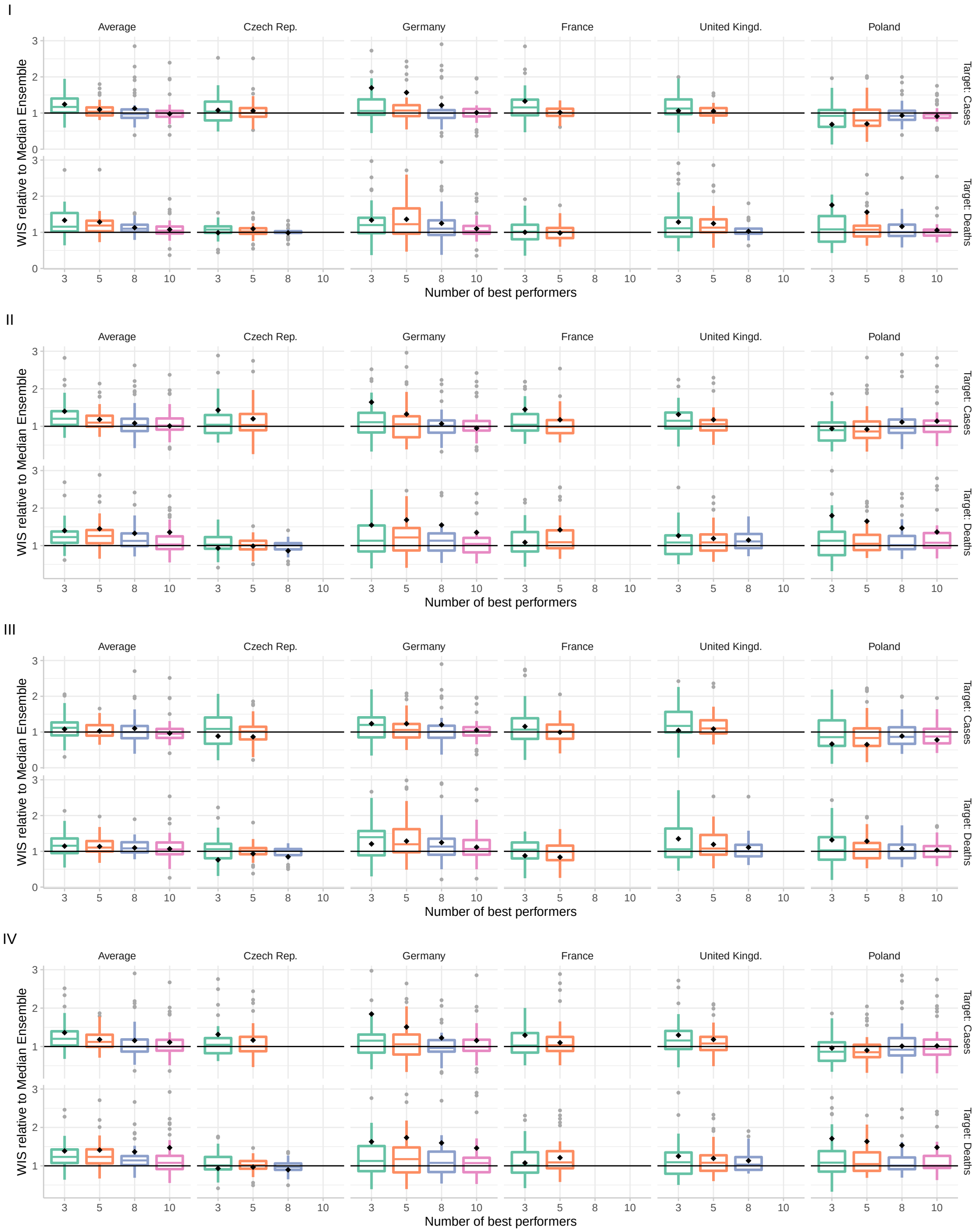


Figure 15: WIS spread of the selection ensemble methods (based on recent component forecast performance) relative to the benchmark (equally weighted median ensemble including all models), by location, target series, and number of best recent performers (k) included. The four panels show the resulting spreads when subsequently applying different ensemble methods to the selected component forecasts. (I): equally weighted median. (II): equally weighted mean. (III): inverse score weighted median. (IV): inverse score weighted mean. For all methods, selection of component forecasters and (where applicable) estimation of weights for a given forecast origin is based on performance in the past four weeks. Outliers with relative scores over 3 were excluded from the plots for legibility - to compensate for this, the diamond-shaped black points represent the mean relative score value. Values below one mean that the respective method outperformed the benchmark for the given target. Missing boxplots for a given k correspond to the cases where the respective combination of location and target type did not have a sufficient model base to support the analysis.

		Agg. method	Average	Czech R.	Germany	France	U.K.	Poland
Cases	$k = 5$	median - unw.	1.10	1.06	1.57	1.02	1.05	0.70
		median - weighted	1.03	0.87	1.23	0.99	1.09	0.65
		mean - unw.	1.18	1.20	1.33	1.17	1.18	0.92
		mean - weighted	1.18	1.17	1.51	1.10	1.18	0.90
	$k = 10$	median - unw.	0.98	-	1.01	-	-	0.91
		median - weighted	0.97	-	1.05	-	-	0.78
		mean - unw.	1.01	-	0.95	-	-	1.14
		mean - weighted	1.11	-	1.16	-	-	1.02
Deaths	$k = 5$	median - unw.	1.29	1.10	1.36	0.98	1.24	1.56
		median - weighted	1.13	0.93	1.29	0.84	1.19	1.28
		mean - unw.	1.45	0.99	1.68	1.42	1.19	1.65
		mean - weighted	1.41	0.97	1.73	1.21	1.19	1.63
	$k = 10$	median - unw.	1.08	-	1.10	-	-	1.06
		median - weighted	1.07	-	1.12	-	-	1.03
		mean - unw.	1.36	-	1.35	-	-	1.36
		mean - weighted	1.47	-	1.46	-	-	1.48

Table 5: Average WIS of the (weighted) selection method, divided by the WIS of the benchmark (equally weighted median including all models). The selection method reduces the set of component forecasters in the ensemble to the number of k best recent performers (as judged by relative WIS in the last four weeks) - different aggregation methods are applied to obtain the ensemble (weighted/unweighted mean/median). Scores are computed by averaging over all forecast dates for the respective combination of location and target series. The “Average” location refers to the setting where scores are averaged across all locations. Scores are missing whenever a location’s model base was not large enough to support the analysis.

for Poland, which we will discuss separately further down. This is also supported by the overall relative scores in Table 5: generally (again excluding the cases series for Poland), the approach works better for $k = 5$ for the three countries where we effectively exclude less models, for all ensemble types. In other words, the closer the setting is to an ensemble that indiscriminately includes all component models for a given location and target series, the better, whether or not the models are subsequently weighted. However, we again make note of the fact that our model base is relatively small. It could be the case that there exists a critical threshold of number of models, after which the added marginal benefit of including another model is smaller and benefits from model selection can actually be realized. If this were the case, we could simply still be sitting firmly in the region of “more models are always better”.

Furthermore, we see that the mean approaches, whether weighted or not, generally see a higher relative average score, a further indication of the general vulnerability of the mean ensemble to outlying forecasts and thereby of the benefit of using the more robust median ensemble, especially when working with a small number of component forecasts. In fact, while additional weighting seems to either not impact or improve the median ensemble, it sometimes has large negative effect on the average score of the mean ensemble, for instance for the cases series in Germany for $k = 5$, which appears to stem from large outlying relative scores (Figure 15). This is likely due to the fact that the mean ensemble is already more vulnerable to deteriorations in performance from one of its member forecasts, especially in small model sets - if such a forecast is additionally given a higher weight due to previous good performance, we can expect performance to further decline in such a situation, thereby driving up average scores.

Thus, since both the weighted and unweighted mean tend to compare unfavorably, we generally focus our remaining discussion on the respective median approaches, specifically the weighted median.

The only case where we see a slight improvement of scores for the average location relative to the benchmark is for forecasting cases and $k = 10$ - the individual results however show that this is only due to the results from Poland, where performance is actually improved across all k, especially for the weighted median ensemble. We now investigate why this is the case, exemplary for $k = 10$, the inverse score weighted median, and in contrast to the cases series from Germany.

Thus, consider Figures 16 and 17, where we display the WIS of the weighted median relative to the benchmark, the 2-week ahead predictions of component forecasts, component forecasts' relative WIS and the assigned weights of the inverse score weighting scheme, over time, and for Poland and Germany, respectively. In these plots, we highlight the five component forecasts that were chosen most often for the respective location.

Figure 16 for Germany neatly shows one of the fundamental issues of selecting and/or weighting models based on recent performance, which is the potential non-stationarity of model's relative performance, and hence the potential unreliability of using past performance as a predictor for current performance. We will now walk the reader through these plots to illustrate that issue.[17] Throughout the summer months of 2021, the weighted method performed comparatively to the benchmark, as it could identify component forecasts that - to a certain extent - showed similarly good performance relative to others. As autumn approaches, it has placed close to half of its weight on four of these component forecasts. It is at this time that cases start rising and (especially at longer horizons), these models overshoot the target. Critically, they overshoot the target more than alternative models: while other models tend to exhibit similar behavior and thus also compare unfavorably to the baseline model of "no change", they still stay closer to the target. The ensemble would thus have fared better if it had drawn from its entire model base and not placed undue weight on a smaller number of component forecasts - as a consequence, the benchmark (which includes all models) performs better than the weighted method during this time. Thereafter, during the month of September, most of these models are sequentially dropped from the set of best performers and hence given zero weight, but only after they have negatively influenced the performance of the ensemble. Moreover, we can see the same behavior again during December of 2021, where models that have received higher weight subsequently overshoot the target and hence undermine the ensemble's performance - the most notable example in this regard is perhaps `FIAS_FZJ-Epi1Ger`, which receives especially high relative WIS during these moments of "collective overshooting" and is consequentially most often dropped altogether from the selected ensemble.

Thus, while the method stays somewhat close to benchmark performance when averaging over all forecast dates (at a relative score of 1.05), it becomes evident that, at least for this series, it undesirably fails in critical times, making it altogether an unattractive alternative to the benchmark.

In contrast to this, we now discuss the cases series for Poland, where the weighted selection ensemble performs substantially better in the aggregate. At the beginning of the study period, the selection ensemble places a large proportion of its total weight on three performers (`MOCOS-agent1`, `ILM-EKF` and `epiforecasts-EpiNow2`) that perform relatively well at predicting the decline in case numbers. As a consequence, the weighted selection ensemble compares quite favorably to the benchmark. Subsequently, during the summer and fall months, model performance is overall more similar and relative performance stays mostly close to one, with some variations. But perhaps most importantly, heading into the winter months of 2021, where Poland (similar to Germany and some other countries in the set) experiences a

[17] At this point, we want to again explicitly mention an important characteristic of the way forecasts are scored here, to clarify the way these plots need to be interpreted. We score models by the forecast date, not by the target date. That is, if a model issues a 2-week forecast on May 3 for a target that realizes on May 15, the agreement of the forecast and the target will be reflected on its score on May 3. Thus, if models for instance severely overshoot the target for a given horizon, this poor performance will be reflected in the scores *before* the target date.

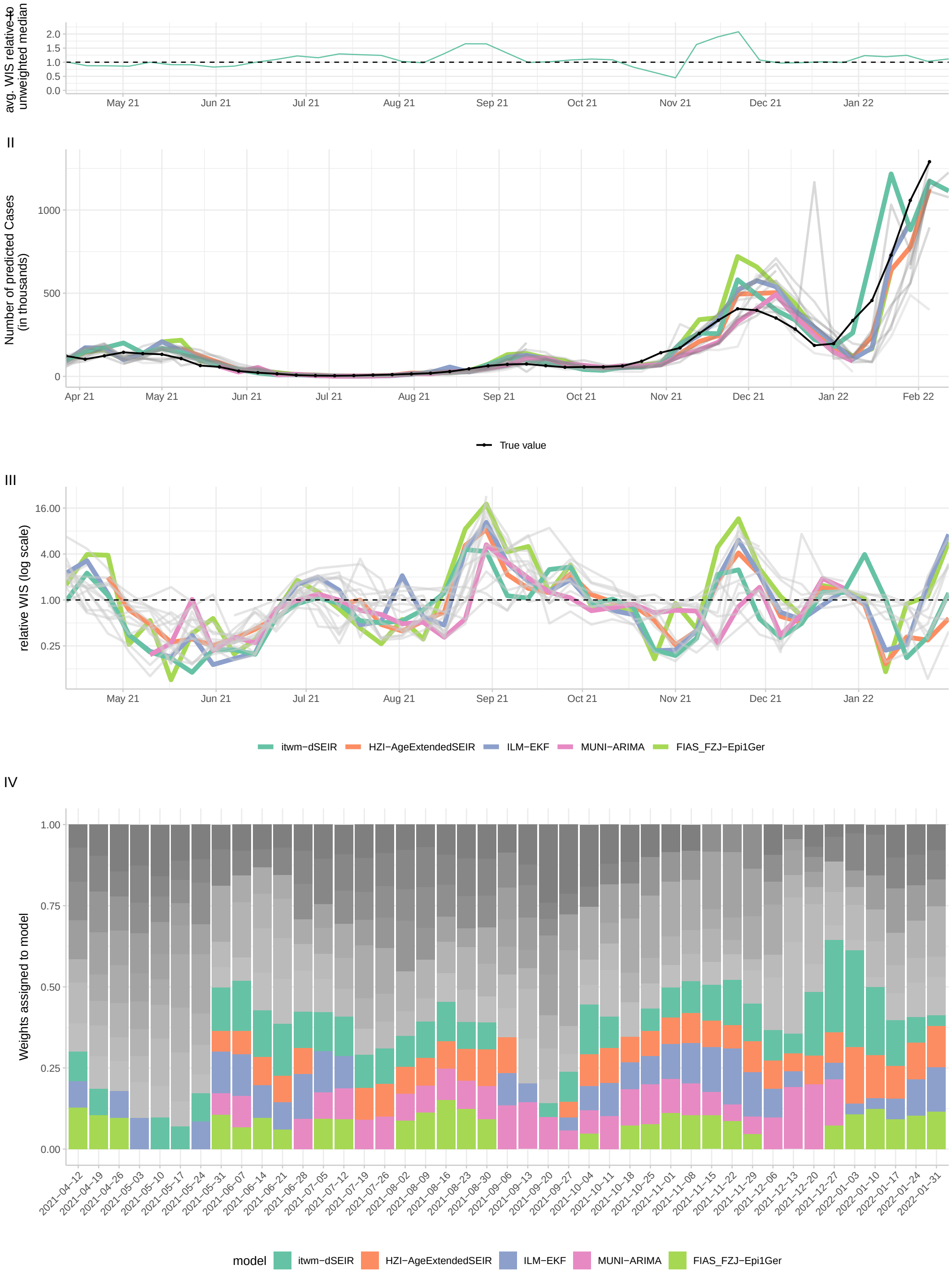


Figure 16: Performance of weekly case forecasts for Germany, for the weighted median selection ensemble based on the ten component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the five models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). The relative performance shown is calculated as a moving average of three weeks, symmetric around the forecast date. Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Plot inspired by Ray et al. (2022).

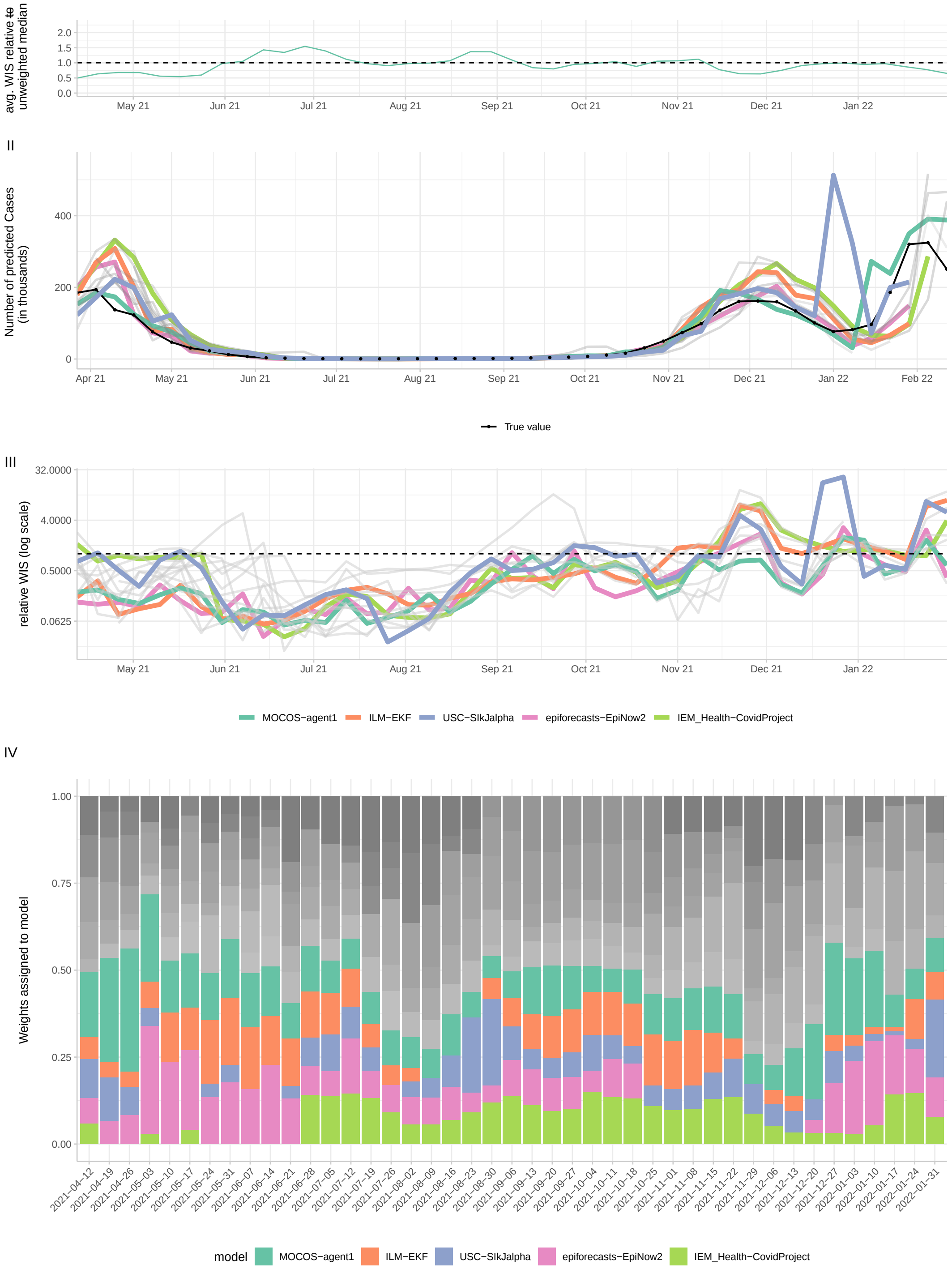


Figure 17: Performance of weekly case forecasts for Poland, for the weighted median selection ensemble based on the ten component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the five models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Furthermore, note that the y-axis of panel (I) was deliberately stretched for compatibility with panel (I) in Figure 16. Plot inspired by Ray et al. (2022).

sharp rise in cases followed by a decline and subsequently another sharp rise, it manages to more heavily weight some models that exhibit relatively less overshooting behavior, thereby overall performing better than the benchmark. One could argue that the behavior near these critical times is actually more meaningful than that during the summer months, which overall have very low incidence rates and thus tend to not be of the same level of interest to decision makers.
In comparison to the series for Germany and concerning the overall distribution of weights, the weights given to the highlighted models in Poland individually as well as overall tend to be higher than those in Germany, indicating that the method is more successful at picking consistent good performers in this location - however, we must also note that this does not yield improved performance during the entirety of the study period. To conclude the discussion for Poland, we would thus argue that Poland contains models (most notably the agent-based `MOCOS-agent1`, which only predicts in Poland) that seem to exhibit particular foresight when forecasting cases - consequently, an ensemble that relies more on these models can see improved performance.
However, there is of course no guarantee that this behavior could be replicated for the same series during other critical times. Furthermore, for the other countries in the set, we also on the one hand don't tend to observe substantially improved performance relative to the benchmark and can at times observe similar behavior to that exhibited by Germany, although trends are often not as neatly aligned and specifically investigation into peak behavior of component forecasts relative to one another towards the end of the study period is complicated by the fact that the number of available models tends to be low (Figures A4, A5, A6). Nevertheless, we would at any rate argue that a central problem with the selection methods is that, as a model might exhibit good performance over a given time period and is thus given nonzero or higher weight in the ensemble, the selection ensemble increasingly makes itself vulnerable to performance fluctuations of such a model.

Specifically for the cases series in the U.S. Hub, Ray et al. (2022) also state that they have observed poor performance of the method during peak times: they give the explanation that the best performing component models for forecasting cases were generally those that were observed to extrapolate recent trends. This would mean that, as trends change from mostly level to increasing, these models would actually cease to be the "best" models (as extrapolating recent trends is fundamentally also associated with overshooting peaks). This is in and of itself an undesirable behavior, but in addition (due to the innately higher level of the series during a peak) also gives higher absolute scores to those forecasts as a consequence. However, precisely as these changes in trends happens, the weighted selection ensemble has already selected for and placed more weight on models that performed well in recent weeks, thereby making itself more vulnerable to their behavior and excluding models which would have mitigated the trend-extrapolating (i.e., overshooting) behavior. By the time it has adapted to this and accordingly downweighted the overshooting models, the peak will naturally have already passed. As stated, we cannot claim that this is the sole and root cause of the selection ensemble tending to perform worse than the benchmark for forecasting cases in our dataset, but given the evidence from Germany, would state that it is at least to a certain degree an issue.

Pertaining to the deaths series, our results somewhat diverge from those in the U.S. Hub, where the series sees consistently improved performance from weighted selection. In our case, in the aggregate, weighted selection actually performs slightly worse for forecasting deaths than for forecasting cases, although we again must note that the somewhat diverging results across locations and therefore would hesitate to call the aggregate difference between the two target series meaningful. Nevertheless, we can definitely not claim a consistent improvement as in the U.S. Hub.

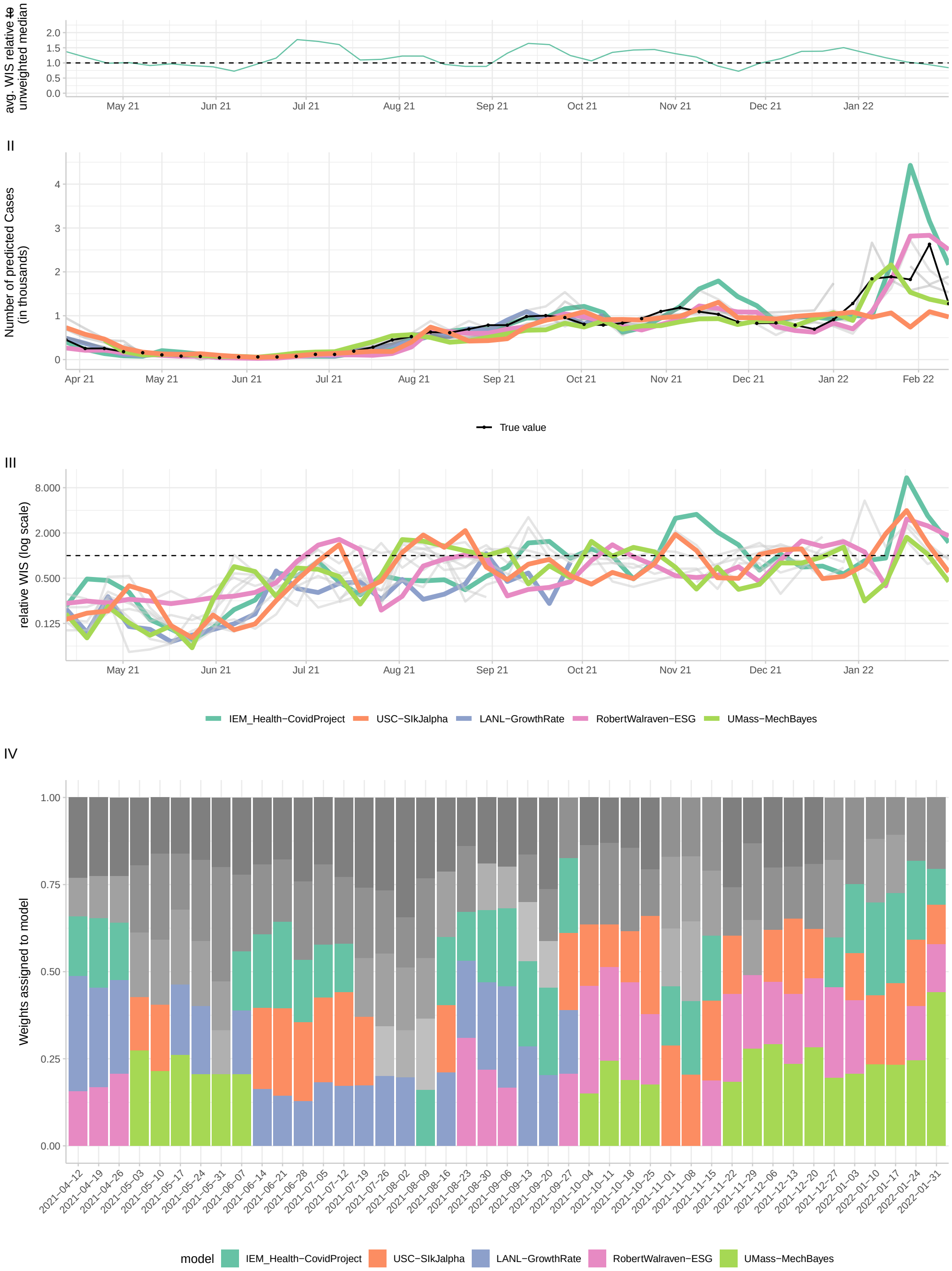


Figure 18: Performance of weekly death forecasts for United Kingdom, for the weighted median selection ensemble based on the five component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the five models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Furthermore, note that the y-axis of panel (I) was deliberately stretched for compatibility with panel (I) in Figure 16. Plot inspired by Ray et al. (2022).

As an exemplary illustration for the deaths series, we now perform the same analysis for the United Kingdom and $k = 5$ in Figure 18. In panel (III), we see that overall, most models seem to be similarly good at forecasting deaths in this location - in fact, over- and under-prediction of the target (and thereby fluctuations in relative performance) seem to be happening at random (i.e., non-predictable) times, which leaves the ensemble usually starting to put weight on a component forecast right as its relative performance starts to deteriorate again. This is also reflected in the assigned weights, with models usually not staying in the set for prolonged periods of time. While we thus don't observe performance that is ever as detrimental as that which we described for the cases series in Germany near peak times, the lack of consistent performance in any component forecasts and the thereby induced failure of the selection method to pick out forecasters at the precise times where they are actually performing well means that the weighted selection method accrues continually slightly higher scores over the entire study period, thereby receiving a poor score compared to the benchmark when averaged over all forecast dates.
As we discussed in section 4.2, deaths are widely considered the easier target to forecast. On the flip side, we would argue that if all models show about the same level of skill at predicting future death counts, this also means that it is harder for any model to perform relatively well to another, and therefore harder for the selection method to pick out meaningful better performers. As Ray et al. (2022) also argue, the trained method likely works well for forecasting deaths in the U.S. Hub simply because there exist forecasters that have a reliable and consistent track record for forecasting that series, while forecast performance for cases tends to be less stationary for all models. Thus, after investigating this for the European Hub, we would argue that such models generally don't seem to exist - while we do sometimes see improved performance for some locations when averaging over the study period, it appears that this is more due the method happening to place higher/lower weight on the "correct" model during periods of higher incidence, and generally not due to an increased level of skill throughout the entire study period (again with a possible exception, France; Figures A7, A8, A9, France: A10).

For the cases series, we also considered the possibility to focus our analysis (both estimation of weights and the evaluation) on shorter horizons, namely for up to two weeks. This is due to the reason that (as mentioned in section 4.2) it is a recurrent result across the Hubs that case forecasts tend to be more unreliable for longer horizons anyway and it thus might be desirable to optimize for the one and two week horizons, which are thereby of more interest to users of these forecasts. However, we observed no clear trend that suggested the (weighted) selection method as it was implemented performs better for shorter horizons (Figure A3) - it appears to work equally well (or rather, poorly) across all four weeks into the future, albeit with some variations. Moreover, considering the estimation of weights, we are already including seven shorter term forecasts vs. three longer term forecasts, so we would not expect any big changes in weights by further restricting this to only horizons one and two weeks ahead. We thus refrained from additionally implementing this strategy.

So far, we have discussed differences in performance only as it relates to the WIS. However, as also mentioned in our introduction of the methodology, it is generally important to also consider other factors of a forecast's performance. We thus now turn to an assessment of the discussed methods' calibration.
To assess the methods' probabilistic calibration, we depict coverage of the central prediction intervals as well as one-sided quantile coverage, separately for the different methods and the benchmark, in Figure 19. For coverage of the central prediction intervals, panel (I) shows that for forecasting cases, all methods are generally overconfident, that is, the predictive distributions' central intervals overall do not cover as many observations as they should, given the respective nominal coverage level. As mentioned before, this is a common result for case forecasts. Furthermore, panel (II) suggests that the ensembles generally

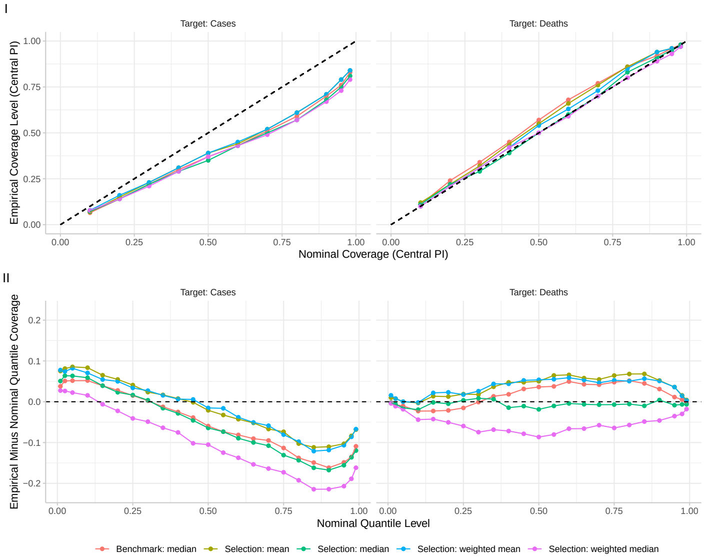


Figure 19: Assessment of probabilistic calibration of the selection ensemble methods based on recent component forecast performance, compared to the benchmark (equally weighted median ensemble including all models). The upper panel (I) shows central interval coverage of the methods, that is, the proportion of observations falling into the predictive distributions' central prediction intervals. The lower panel (II) shows the difference between the empirical and nominal coverage of the predictive quantiles. Negative (positive) values indicate that less (more) observations fell below a predictive distribution's respective quantile than required. For both panels, the black dashed line corresponds to optimal coverage levels. For the selection methods, results are based on selecting five best performers for locations with comparatively small model base (Czech Republic, France, United Kingdom) and ten best performers for locations with comparatively large model base (Poland, Germany).

have more of an issue with the upper rather than the lower quantiles, which indicates that the predictive distributions have some downward bias - this is however more the case for the median-based approaches than the mean-based ones, with the weighted median showing most downward bias out of the methods considered. These results are mostly in line with the results of Ray et al. (2022) for the European Hub, but in contrast to them, we observe that the weighted mean's quantile coverage is more in the line with that of the unweighted mean than that of the weighted median.

For forecasting deaths, empirical coverage of the central prediction intervals is in general closer to nominal coverage, with only slight under-confidence for the benchmark and the mean-based ensembles. Again, close to nominal central interval coverage rates are a common result for ensemble forecasts for deaths. According to panel (II), the weighted selection median is slightly downward biased, with consistently less observations falling below the respective quantiles than required - this is also reflected in its bias measure (which sits at -0.123). In contrast to this, the mean-based approaches show some slighter upward bias, while the remaining median-based approaches' empirical quantile coverage is more in line with nominal coverage. These results slightly deviate from those of Ray et al. (2022) for the European Hub - in particular, they find that the weighted median is over-confident, but not biased.

Overall, the methods are thus mostly similar in terms of probabilistic calibration. However, we must note that while the weighted median tended to perform best out of the selection methods considered with respect to relative WIS, it compares less favorably to other methods when assessing calibration, mostly due to its downward bias.

Before we close this section, we discuss some more interesting points that emerged from this analysis.

An interesting lead emerges when consulting Table 5. For the case of $k = 5$, we can see that the weighted

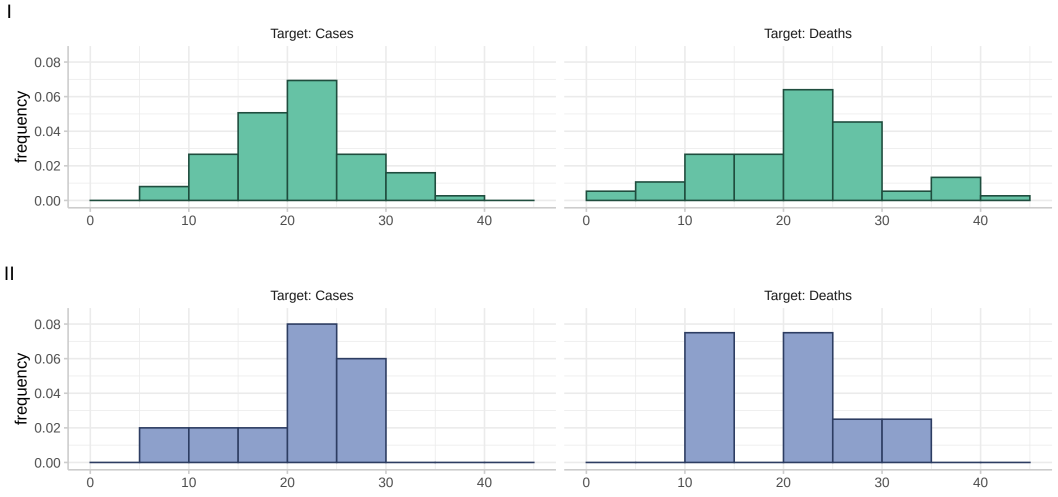


Figure 20: Distribution of the number of times individual models were chosen by the selection ensemble for the 43 weeks considered, based on the chosen sets of $k = 5$ for countries with smaller model base (Czech Republic, France and the United Kingdom) and $k = 10$ for countries with larger model base (Germany and Poland), after accounting for confounding factors (number of available models at each forecast date and individual model availability). Models which were available for less than 15% of the study period were excluded, to not skew results. Panel (I) shows the resulting distributions based on all models, while panel (II) only considers models that issued forecasts for a single location. In total, there were 76 unique combinations of model and location for forecasting cases (with 10 models forecasting for a single location) and 77 for forecasting deaths (with 8 models forecasting for a single location).

median mostly performs better than the unweighted median for both target series, suggesting that in very small model sets, it could actually be of benefit to weight models by their recent performance, rather than a source of extra noise. Put differently, while it does not emerge as a dominant strategy to actively *reduce* the set of models, weighting could be applied to the median ensemble in situations where only a small number of models is available to begin with. In a future analysis, this could be further investigated via other locations in the European Hub, which overall tend to have a lower number of available models than the locations considered here.

Furthermore, we thought it interesting to further analyze the number of times individual models were chosen by the selection ensemble for the two target series and thereby investigate if models tended to be chosen equally often or if the resulting distributions tended to be more spread out. This would give an indication on whether individual models tended to perform equally well or if there existed models that showed higher forecasting skill over time. On the one hand, this might yield further understanding for why the selection ensemble does not tend to lead to improved performance, but we think it is also interesting in and of itself to judge whether consistent performance differences tend to arise between component forecasts. It thus might be possible to establish that consistently “good” or “bad” performers exist - for instance, if a model was never in the set of best performers as judged by relative recent skill over the 43 forecast dates considered, we could safely call such a model a poor performer.
Thus, the more concentrated these distributions are, the more uniform we can think forecast skill to have been over the study period, while more spread out distributions would indicate that there existed models that overall had higher relative skill compared to others. We also consider whether distributions differ depending on whether all models or only those that forecast for a single location are regarded. As stated before, it is conceivable that these “single-location” models might show greater skill compared to

models that aim to forecast more broadly, presumably as the former can be tuned more heavily to the circumstances and characteristics of a specific location.

For each combination of target series and location, we thus summarized the number of times each model was chosen by the selection method, after accounting for confounding factors.[18] There were only three examples of models with sufficient availability that were never chosen for a combination of location and target series,[19] showing that most models had at least intermittent periods or single dates where they performed relatively well.

The resulting distributions for both overall models and the single-location models are shown in Figure 20. Considering the distribution of all models in panel (I), differences between the target series are not that pronounced, although the distribution for the deaths series does appear to be slightly more spread out and negatively skewed. This suggests that performance differences between models are slightly more pronounced for the deaths series - in particular, there does seem to be a small number models that were chosen more consistently. Conversely, the distribution for cases appears to be more concentrated, indicating that performance was more equal overall. Considering the distribution of the single-location models in panel (II), it appears that for cases, these models did tend to be chosen slightly more often, indicating that tuning a model to a specific location might indeed be associated with higher relative skill for forecasting cases. Conversely, we tend to see the opposite holding true for the deaths series. We must however again note that this comparison is based on a small number of single-location models, most of which forecast for Germany or Poland.[20] It would thus be interesting to repeat a similar analysis for other locations in the European Hub, a larger number of which has modeling teams which only submit forecasts for the respective location. If it becomes evident that such models do tend to perform better, this might in turn also increase the viability of a weighted approach for these locations.

Importantly, these distributions also give an indication of why the approach might not perform so well overall: the fact that the distributions are generally more concentrated around the value 20 (roughly half of the study period) once more indicates that models are often switched out and we in particular don't seem to have many models that emerge as *consistently* good performers.

Overall, we argue that we succeeded in giving some explanations why the (weighted) selection method might not consistently work in our given setting. First and foremost, it seems that for most of the combination of locations and target series considered, there do not exist models which exhibit consistently good performance relative to other models for extended periods of time. Due to this fact, the assigned weights as well as the included models in the selection ensemble often fluctuate. Fundamentally, due to the lagged nature of the weight estimation and inclusion/exclusion decisions, models are given higher weight due to *past* signals, but subsequently might exhibit drops in performance. This behavior thus gives rise to fluctuations in the performance of the selection ensemble relative to the benchmark.

An additional stumbling block for the method can be changes in trends of the underlying series, especially if such a structural break is also associated with changes in relative skill of component forecasts.

[18] Confounding factors were the number of available models at each forecast date (for e.g. $k = 5$, being chosen as one out of 6 models is less meaningful than being chosen as one out of 10), as well as a model's individual availability over the entire study period. A model's binary "chosen"-status at each forecast date was thus multiplied by a factor a_t ($a_t = \frac{M_t}{k}$, where M_t is the number of available models), and these scaled status were summed over the study period to receive a model's total count, which was subsequently inversely scaled by its overall availability. Forecast dates for which $k = M_t$ for the given combination of location and target series were excluded, as well as models which were available for less than 15% of the study period for a combination of location and target series, to not skew results.

[19] `MUNI-VAR` for the Czech Republic (both target series) and `UMASS-SemiMech` for U.K. deaths, but note that these models were chosen for other locations and series.

[20] 7 of the 10 single-location models for cases and 7 out of the 8 single-location models for deaths are from either Germany or Poland.

Particularly if models that were given higher weight due to previous low relative scores overshoot the target after a change in trend, this can lead to overall poor performance of the selection ensemble. On the one hand, these are the periods that will most drive aggregate scores, and they are also the periods that are of the most interest to decision makers. During these times, it would often be beneficial for the ensemble to draw on its entire model base, to mitigate the effects of those models that overshoot the target.

A general exception could be a situation such as the cases series for Poland, where there do seem to exist some models that exhibit better foresight in predicting future case numbers for the method to successfully select and place weight on, but we would still be cautious in recommending such an approach without further investigation into (peak) behavior in following times.

Ultimately, our advice would thus be that one needs to consider the respective model set and decide whether forecasters with good consistent foresight exist. We would thus argue that in lieu of the existence of component forecasts that actually perform reliably well for an extended period of time, (weighted) selection does not seem to be a viable strategy. Usually, individual model performance is not stationary enough, again highlighting the benefits of simply using an equally weighted median ensemble that includes all models. This simple strategy, after all, has the precise benefit of being able to better mitigate occasional deteriorations in performance from one or some of its member models and not being reliant on measures of past performance.

In the following section, we investigate whether weighting based on performance shown by general modeling strategies, rather than based on single component forecast performance, can improve ensemble predictions.

6.2 Weighting based on modeling strategies

In this section, we discuss our efforts of implementing an ensemble composition scheme that is based on the modeling strategies as discussed in section 2.1.

The idea of selecting component forecasts for an ensemble based on their modeling strategy does have some precedent in the literature. In their analysis of the deaths series in the U.S. Hub, Taylor and Taylor (2021) identified that for states that had relatively low overall mortality compared with others, ensembles that exclusively consisted of mechanistic models gave the most accurate forecasts, while the same strategy gave similar performance for states with medium mortality and worsened performance for states with high mortality, compared to an ensemble including all types of models. Apart from this, we are unaware of any studies attempting to base ensemble composition or weights on model types.

We thought about implementing a strategy similar to that of Taylor and Taylor (2021), but given the small number of locations in our set, which furthermore show similar levels of incidence when accounting for population size, we did not think this to be a viable idea. Alternatively, we considered the idea of categorizing phases of the pandemic (based e.g. on some notion of “stable”, “upswing”, “downswing”), as we also believed that differences between modeling strategies would (at least conceptually) show along these lines - we however also ultimately found that any such categorization would be arbitrary to a non-negligible degree, since most series exhibited intermittent periods of growth even during the summer months (which did see lower incidence in general).

Concerning our results in section 5.1, it was not possible to give an overall consistent ranking of model types in the data. One could however surmise that in specific situations, one general type of modeling philosophy could be better suited than another for giving accurate predictions. We also saw some tentative evidence of this: for most locations, the group of semi-mechanistic models received critically high scores due to overshooting during the last period under study, for both target series. Consequently, it might have been beneficial for most locations to accordingly downweight this group and rely more heavily on the forecasts made by statistical or mechanistic models. However, since the goal is to mimic the real-time situation as closely as possible, implementing an ensemble that is directly based on this information would rely too much on using hindsight knowledge. Hence, given our findings that selecting and/or weighting *individual* component forecasts based on recent performance measures does not seem to reliably improve the ensemble, the question arose whether it might be possible to leverage differences in recent performance at the level of modeling strategies. The natural choice was thus to implement a weighting scheme that estimated weights based on these modeling strategies.

Before we describe our implementation and the results, we discuss some reasons why we thought this could work, as well as some reasons why it might not. As already mentioned, we did observe some tentative evidence that some modeling strategies performed better than others during some periods, and we wanted to test whether an automatic weighting scheme might be able to pick up on this *and* consequently use it to improve ensemble performance. Moreover, grouping models and thereby weighting at the level of model types rather than individual models has the advantage of being able to use all models, even if they newly enter into the Hub and no individual record of performance is available yet. Furthermore, due to models dropping in and out of predicting, there sometimes exist gaps for each model, so we often lack a consistent record of recent performance to base weight estimation on. On the level of model types, we don't have these gaps.[21] Lastly, this procedure presumably removes some variability from the practice of estimating weights: if a single model has some fluctuations in performance, this will impact a grouped score less than individual scores.

However, while model types do share a common approach to modeling, they are of course not a monolith and characteristics other than their categorization within these groups can dominate their performance. The categorization might thus not be meaningful enough. Importantly, this strategy could presumably suffer from the same problem as that which we discussed in the previous section: performance could not be stationary enough to reasonably allow for estimation of weights and a weighted ensemble might thus be too slow at adapting to changes in performance from the model groups.

Due to the lower availability of the other modeling strategies, we again only focus on models from the categories mechanistic, semi-mechanistic and statistical. We thus devised two different methods: for the first, we decided to use the inverse score weighting approach as used in the previous section, but computed the average score in the respective window at the level of modeling strategies rather than single models. The resulting weights were then divided by the number of models available for the given modeling strategy and assigned to each member model. In other words, this amounts to weighting the groups unequally, but giving each component forecast within each group equal weight. Differently to before, we decided to use an exponential weighting scheme as described in Section 3.3 (with $\alpha = 0.5$) to calculate the recent score of model types, with the hope that this would let the method adapt quicker to changes in performance from the model groups, while simultaneously still including faraway scores so as not to increase the variance of the estimation too much, albeit to a lesser degree.

As a second method, we first built an equally weighted median ensemble from each of the model types, thus in effect giving three smaller ensemble models. For these, we then calculated QRA weights as detailed in section 3.3. We however noticed that the resulting weights could be quite variable and also sometimes tended toward the extreme ends (in some weeks exclusively giving all weight to one of the

[21] except for a brief section at the beginning of the study period where we have no statistical models the Czech Republic, France and the United Kingdom for forecasting cases.

	Method	Average	Czech R.	Germany	France	U.K.	Poland
Cases	Inv. score w. median	0.983	1.021	1.152	0.989	0.895	0.932
	QRA	1.240	1.239	1.085	1.177	1.488	0.972
	QRA - adj. weights	1.145	1.165	1.006	1.119	1.287	1.020
Deaths	Inv. score w. median	1.060	1.059	1.054	0.856	1.047	1.214
	QRA	1.336	1.124	1.240	1.085	1.172	1.765
	QRA - adj. weights	1.188	0.948	1.096	1.191	1.021	1.453

Table 6: Average relative scores of ensemble forecasts that were composed by calculating weights at the level of the dominant modeling strategies within the European Hub (mechanistic, semi-mechanistic, statistical), compared to an equally weighted median based on the same models. The inverse score weighted median ("Inv. score w. median") calculated weights based on the average score each group of models obtained during the last four weeks, via an exponential weighting scheme that gave higher weight to recently obtained scores - the ensemble was built via a weighted median of all component forecasts (calculated weights were equally distributed between models that belonged to a group). The QRA method calculated weights for three ensembles that were built from the model groups and subsequently combined these via a weighted mean. The adjusted QRA method ("QRA - adj. weights") adjusted weights of the QRA method to account for the differing number of models within each group.

ensembles, Figure A12).[22] Furthermore, as we thought it could be wise to additionally account for the differing availability of models for the respective model types over time amount of models within each group (see Figure A11), we additionally decided to implement a simple adjustment strategy that amounts to balancing the weights equally between those that the QRA algorithm calculated and those that would result if weighting the ensembles purely based on their size.

Since we want to focus on the general viability of the approach, we here use the equally weighted median ensemble based on the same set of models as a benchmark - in practice, this thus excludes the expert judgment based models as well as the two agent-based models for Poland. This allows us to judge whether a weighting via categorization of modeling strategies is *in principle* fruitful or possible, as we remove the confounding factor of ensemble size.

The results, again stratified by location and target type, are shown in Table A2. Similar to before, no clear general improvement upon the equally weighted median can be observed for any method. The inverse score based median ensemble compares similarly to the benchmark, for both cases and deaths, although this again varies somewhat by location. The QRA ensemble tends to performs the worst - we expect that this is due to the fact that it is essentially a small mean ensemble and thus more vulnerable to fluctuations in performance from one of the three ensembles it includes, which is likely again exacerbated by the more extreme distribution of weights (Figure A12). The QRA ensemble based on the simply adjusted weights mostly tends to improve on the QRA ensemble, supporting the claim that the weights given by the pure QRA method were indeed too extreme at times.

For full disclosure, we also calculated results in comparison with the *full* median ensemble (Table A2). Relative scores tend to be slightly worse, but overall very similar for both the aggregate and across locations, with the exception of Poland (especially for the cases series). This is likely due to the fact that we are excluding two extra models for this location, one of which (`MOCOS-agent1`) the previous section has additionally shown to be a particular good model for forecasting cases. In terms of the methods' calibration (Figure A13), coverage rates are overall similar to before, notably with the weighted median showing downward bias, and a general tendency for over-confidence for predicting cases.

Since the inverse score weighted median again tended to perform the best, we focus our remaining discussion on this method. Panel (I) in Figure 21 shows the average relative performance to the benchmark,

[22] The method's tendency for sparsity was also noted by Ray et al. (2021).

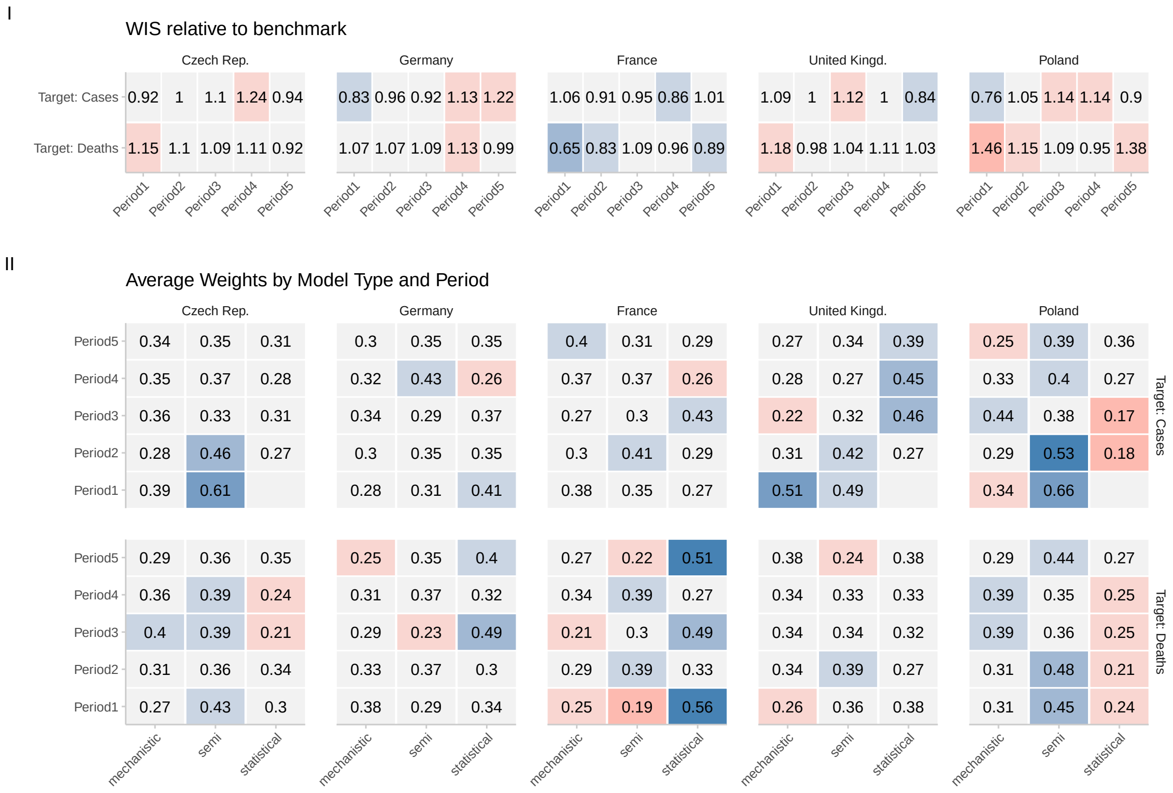


Figure 21: Panel (I): Average WIS of the weighted median ensemble (weights calculated by inverse score weighting at the level of modeling strategies), scaled by the WIS of an equally weighted median ensemble ("benchmark" model) based on the same set of component forecasts. Scores are calculated by location, target series, period and horizon. Relative scores above 1 correspond to the weighted median performing worse than the benchmark, while values below 1 correspond to it performing better than the benchmark. Panel (II): Average weights given to each modeling strategy during a given period, by location and target series. The code to produce these plots was adapted from the `scoringutils` package (Bosse et al., 2022b).

separately by period. A central determining factor for the aggregate scores in Table A2 is once more the performance during high-incidence times: for instance, while the method works similarly well as the benchmark for Poland during the middle periods for forecasting deaths, larger scores in periods 1 and 5 drive up the aggregate score. Likewise, the good relative performance in the UK for forecasting cases seems to stem from period 5, where it receives a much lower score.

Relative performance is mostly similar across horizons, although exceptions in both directions (better/-worse relative performance with increasing horizon) exist, with no systematic trends becoming apparent (Figure A14).

Concerning the distributions of average weights by period (Panel (II) in Figure 21), these are often quite uniform, suggesting that the method doesn't pick up on meaningful differences between the modeling strategies, at least not for prolonged periods of time. Furthermore, even when weights are more unequally distributed, this is not necessarily associated with improved performance: for instance, higher weights are assigned to statistical models for forecasting cases in the UK in periods 3, 4, and 5, but performance is only improved in the last period. Conversely, the assignment of higher weights to statistical models is associated with improved performance for the deaths series in France.

Furthermore, recall our argument that it might have benefited an ensemble to give less weight to the group of semi-mechanistic models during period 5. The fact that this model group received high relative scores is mostly not reflected in the assigned average weights in period 5, although they tend to be downweighted towards the end of the period in most locations (see right column, Figure A12 for some examples). The reason for this is likely that higher scores were accrued *during* this period, with weights simply not adjusting quick enough, despite the exponential weighting scheme. Even in the two cases where semi-mechanistic models are given lower weight, results are not clear: we see improved performance for forecasting deaths in France, but not for the United Kingdom (although the difference in scores is also not that large there, see Figure 10).

We deliberately keep the discussion for this section shorter, as upon investigation we often found that behavior of the ensemble we built here was often similar to that of the weighted selection median in the previous section: relative to the benchmark, we observed both periods of improved performance, as well as worse performance - sometimes deteriorations in performance happened during or following a peak in the respective series, but also often as a result of seemingly erratic fluctuations in weights and the method thereby giving higher weight to a modeling strategy just as its relative performance to others deteriorated again. Furthermore, just as before, performance relative to the benchmark did not appear to follow any neat patterns across locations, target series or over the time period under study. Lastly, as already stated, whether or not the method did receive lower scores in the aggregate usually solely depended on whether or not it happened to correctly place weights on the better modeling strategy for the respective location and target series during the periods with higher incidence levels.

Fundamentally, the method thus also suffered from what we would identify to be the central problem with weighting based on past performance, as also discussed in the previous section: estimated weights fundamentally lag behind forecaster's real-time performance, and relative performance can change too rapidly to allow for estimation of weights that would actually be beneficial for ensemble composition. We thereby would continue to recommend to rely on an equally weighted median of all models when building an ensemble.

7 Conclusion and discussion

The goal of this thesis was to investigate modeling strategies and ensemble behavior in the European COVID-19 Forecast Hub, with the potential end goal to be able to identify characteristics of both that would allow themselves to be leveraged for ensemble composition and thereby improve the performance of an ensemble in the European Hub.

Investigating the impact of modeling strategies on model performance, the major modeling strategies (mechanistic, semi-mechanistic, statistical) tended to perform similarly in terms of the WIS in the aggregate, while performance differences were sometimes pronounced at lower resolutions. Mostly, these however did not follow neat patterns, with fluctuations becoming apparent along the lines of locations, periods and the two target series. Semi-mechanistic models did tend to perform worse in terms of the WIS than other modeling strategies at longer forecast horizons for forecasting cases, while the assessment of calibration revealed that this group performed consistently similar to or better than others, thereby highlighting the importance of generally regarding different evaluation metrics when assessing forecasts. Due to modeling strategies overall performing similarly, we have given further indication that basing models on epidemiological principles - be it via explicitly modeling transmission dynamics or relying on approaches based on estimating the growth rate or effective reproduction number - does not provide a systematic benefit in accuracy for short-term forecasting of COVID-19, compared to models that are agnostic to these principles. In a further analysis, one could also characterize models based on other criteria, most notably by the information they incorporate into their models, such as whether they explicitly account for non-pharmaceutical interventions, and thus investigate whether these characteristics might lead to improved performance.

With the goal of better understanding the performance of an ensemble with respect to its member models, we devised an experiment where we systematically added models based on relative performance and distance measures to an existing base ensemble, to investigate whether we could observe relative performance differences in the response of an ensemble to new additions in a controlled setting. While the additional models tended to have positive effect on the ensembles in most cases, we also found that it was generally safer (that is, leading to less drastic performance differences on average) to add models that had recently shown better performance or were more in agreement with the established base ensemble. However, while effect sizes consistently emerged, and in particular showed that the median ensemble was more robust to additions than the mean ensemble, it was difficult to directly interpret their magnitude. We think it would be interesting to repeat this analysis while keeping the composition of the ensemble constant over several forecast dates, as interest perhaps typically lies more in longer term additions to an ensemble. Furthermore, it would be interesting to investigate whether effects continue to shrink for even larger ensembles.

In a further step, we aimed to build on the knowledge gained in the previous analyses as well as in related studies by devising and implementing strategies that tune ensemble composition, with the goal of improving ensemble performance. For neither the cases nor the deaths series, we were able to establish methods that consistently performed better than the benchmark equally-weighted median ensemble, neither when preselecting models that showed relatively good performance in recent weeks, nor when basing weight estimation on modeling strategies. In relation to the latter, while our retrospective analysis showed that differences between modeling strategies arose at times, when trying to mimic a real-time setting and simply relying on measures of recent past performance on how to weight model types, we saw that it was not possible to consistently make use of these performance differences. Importantly, this also further suggests that there were usually no consistent performance differences over prolonged periods of time between the different modeling strategies.

In general, we found that the fundamental problem with approaches that aim to base ensemble composition on past performance on (groups of) models is that it composes the model set on a lagged measure, which, given the non-stationarity of model performance, is not a reliable predictor for current performance. Once a model is included in the ensemble and/or given higher weight, potential performance deteriorations of individual models will have a bigger effect than if the ensemble instead drew from the entire available model base equally. The distributions of number of times models were chosen gave further indication of the fact that for both target series, there generally did not exist *consistently* good performers, as models tended to be chosen equally often.

Thus, even though we did not succeed in establishing a promising alternative method that could be employed and evaluated along the current median ensemble in a true real-time analysis, we argue that there is an important advantage to this analysis: even with mild retrospective knowledge, we didn't succeed in devising a strategy that consistently outperformed the median ensemble, thereby supporting both its previous as well as its continued use in the European Hub.

Nevertheless, we think that there are yet some strategies that could be tried in a future analysis: our results from the recombination experiment showed that an ensemble's performance is generally more vulnerable to the inclusion of models that are in disagreement with it. One could thus also implement a strategy that excludes a model from an ensemble if its individual distance from the ensemble is too large.[23] While this is conceptually similar to removing outlying forecasts, it formalizes and automatizes their removal, in contrast to excluding them manually. As a justification of this practice, if we conceptually regard the ensemble as a method to generate consensus, it might be fair to remove models that disagree with the ensemble, as it is not unlikely that such disagreements might have arisen from an error, for instance in the modeling pipeline. Especially when used in conjunction with weighting methods, we think that this could lead to greater robustness, due to cases where the ensemble would otherwise give undue weight to an outlying forecast.

We want to close with a note on our evaluation method. Being aware of the potential problem of locations with higher levels of incidence critically influencing aggregate results, we generally decided to regard locations separately in addition to reporting aggregate results, both for evaluation of alternative ensembles and for the modeling strategies. We thought this method to be more honest - as amply demonstrated, superior performance in the aggregate can emerge from very different base behavior. Nevertheless, this often made for a somewhat ungainly analysis, as results that appeared in the aggregate often showed to stem from diverging results between locations and explaining these differences was usually not possible or risked dwelling too extensively on single phenomena. Overall, we would thus argue that the locations considered might be too heterogeneous to establish universal results, particularly for the recommendation of alternative ensemble strategies. In a similar vein, as aggregate results were often critically influenced by periods with larger levels of incidence, we strove to counteract this by stratifying our reporting of results by period, but this also suffered from the issue of further increasing the dimensionality of the results output and increased the difficulty of determining potential sources of the differences. While we presented arguments stating that the vulnerability to high absolute scores is a meaningful characteristic of aggregation methods, we would argue that this also has the downside of severely limiting comparability of results when evaluating both individual models, model groups, or competing ensembles.

[23] Since distance measures also scale with the target series, this could be formalized by the removal of a model reducing the average pairwise distance of an ensemble by a certain proportion.

Bibliography

Adrian Baddeley and Rolf Turner. spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

Nikos Bosse, Sam Abbott, Hugo Gruson, Sebastian Funk, and Nicholas G Reich. epiforecasts/scoringutils: 1.0.0, May 2022a. https://zenodo.org/record/4618017.

Nikos Bosse et al. Comparing human and model-based forecasts of COVID-19 in Germany and Poland. December 2021. http://medrxiv.org/lookup/doi/10.1101/2021.12.01.21266598.

Nikos Bosse et al. Evaluating Forecasts with scoringutils in R. 2022b. arXiv: 2205.07090, Version Number: 1.

Johannes Bracher, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17(2):e1008618, February 2021a.

Johannes Bracher et al. The German and Polish COVID-19 Forecast Hub, 2020. https://github.com/KITmetricslab/covid19-forecast-hub-de.

Johannes Bracher et al. A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nature Communications*, 12(1):5173, December 2021b.

Fred Brauer and Carlos Castillo-Chavez. Epidemic Models. In *Mathematical Models in Population Biology and Epidemiology*, volume 40, pages 345–409. Springer New York, New York, NY, 2012.

Logan C. Brooks et al. Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the U.S. 2020. (International Institute of Forecasters, 2020).

Gerda Claeskens, Jan R. Magnus, Andrey L. Vasnev, and Wendun Wang. The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762, July 2016.

Estee Y Cramer et al. The United States COVID-19 Forecast Hub dataset. preprint, November 2021. http://medrxiv.org/lookup/doi/10.1101/2021.11.04.21265886.

Estee Y. Cramer et al. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, 2022.

Francis X. Diebold and Roberto S. Mariano. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, July 1995.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, May 2020.

Christophe Fraser. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. *PLoS ONE*, 2(8):e758, August 2007.

Sebastian Funk et al. Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15. *PLOS Computational Biology*, 15(2):e1006785, February 2019.

Sebastian Funk et al. Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. preprint, November 2020. http://medrxiv.org/lookup/doi/10.1101/2020.11.11.20220962.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007.

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

Inga Holmdahl and Caroline Buckee. Wrong but Useful — What Covid-19 Epidemiologic Models Can and Cannot Tell Us. *New England Journal of Medicine*, 383(4):303–305, July 2020.

Ruth Ann Jajosky and Samuel L Groseclose. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *BMC Public Health*, 4(1):29, December 2004.

Lyndon P. James, Joshua A. Salomon, Caroline O. Buckee, and Nicolas A. Menzies. The Use and Misuse of Mathematical Modeling for Infectious Disease Policymaking: Lessons for the COVID-19 Pandemic. *Medical Decision Making*, 41(4):379–385, May 2021.

Michael A. Johansson et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274, November 2019.

T. N. Krishnamurti et al. Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, 13(23):4196 – 4216, 2000.

Craig J. McGowan et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*, 9(1):683, December 2019.

Jessica E. Metcalf and Justin Lessler. Opportunities and challenges in modeling emerging infectious diseases. *Science*, 357(6347):149–152, July 2017.

Kelly R. Moran et al. Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast. *Journal of Infectious Diseases*, 214(suppl 4):S404–S408, December 2016.

Jakub Nowotarski and Rafał Weron. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, 30(3):791–803, September 2015.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Evan L. Ray et al. Challenges in training ensembles to forecast COVID-19 cases and deaths in the United States. 2021. (International Institute of Forecasters, 2021).

Evan L. Ray et al. Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. 2022. arXiv: 2201.12387 Version Number: 2.

Nicholas G. Reich, Ryan J. Tibshirani, Evan L. Ray, and Roni Rosenfeld. On the predictability of COVID-19. *International Institute of Forecasters*, 2021. (International Institute of Forecasters, 2021).

Nicholas G. Reich et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019.

Nicholas G. Reich et al. Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. *American Journal of Public Health*, 112(6):839–842, June 2022.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554, 2005.

Katharine Sherratt et al. European Covid-19 Forecast Hub, August 2022a. https://doi.org/10.5281/zenodo.6962430.

Katharine Sherratt et al. Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations, 2022b. https://www.medrxiv.org/content/10.1101/2022.06.16.22276024v1.full.pdf.

James W. Taylor and Kathryn S. Taylor. Combining probabilistic forecasts of COVID-19 mortality in the United States. *European Journal of Operational Research*, June 2021.

Thordis L. Thorarinsdottir, Tilmann Gneiting, and Nadine Gissibl. Using Proper Divergence Functions to Evaluate Climate Models. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):522–534, January 2013.

Ryan Tibshirani and Logan Brooks. quantgen: Tools for generalized quantile modeling, 2020. https://github.com/ryantibs/quantgen.

Allan Timmermann. Chapter 4 Forecast Combinations. In *Handbook of Economic Forecasting*, volume 1, pages 135–196. Elsevier, 2006.

Yijin Serena Wang et al. covidHubUtils: Utility functions for the COVID-19 forecast hub, 2022. https://github.com/reichlab/covidHubUtils.

Lila Warszawski et al. The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework. *Proceedings of the National Academy of Sciences*, 111(9):3228–3232, March 2014.

Teresa K. Yamana, Sasikiran Kandula, and Jeffrey Shaman. Superensemble forecasts of dengue outbreaks. *Journal of The Royal Society Interface*, 13(123):20160410, October 2016.

Jon Zelner, Julien Riou, Ruth Etzioni, and Andrew Gelman. Accounting for uncertainty during a pandemic. *Patterns*, 2(8):100310, August 2021.

A Appendix

	μ		σ	
	Median Ens.	Mean Ens.	Median Ens.	Mean Ens.
β_0	1.0257	1.0075	-1.3339	-1.5743
	0.0013	0.0009	0.0030	0.0031
Best Model	-0.0145	-0.0230	-0.0091	-0.1039
	0.0013	0.0009	0.0030	0.0030
Worst Model	0.0193	0.0301	0.0222	0.1523
	0.0013	0.0011	0.0030	0.0030
Max. Distance Model	0.0234	0.0550	0.0949	0.4347
	0.0014	0.0014	0.0030	0.0030
Min. Distance Model	-0.0148	-0.0159	-0.1896	-0.7245
	0.0012	0.0008	0.0030	0.0030
Horizon	0.0027	0.0010	0.0648	0.0542
	0.0004	0.0002	0.0009	0.0009

Table A1: Coefficients from a fitted GLM with gamma-distributed outcome. Standard errors are reported below coefficients. The outcome variable is the WIS after adding different types of models from a proposed set of four models to base ensembles, relative to the WIS of the base ensemble before the addition. Base ensembles consisted of four member models, and effects were assessed separately depending on the type of base ensemble (quantile-wise median or mean ensemble). Models were added based on the characteristics: best/worst performer, minimum/maximum distance to the base ensemble and a reference (randomly chosen) model. Dummy effects were added for the type of model, with the random model as base category. Forecast horizon was added as control, and random effects were included to control for location, target type and period. For this specification, outcome values above 3 were excluded. Model was fit using the `gamlss` package (Rigby and Stasinopoulos, 2005) - the μ parameter corresponds to the mean of the response distribution, σ^2 to its dispersion. Importantly: $N = 266497$, size of standard errors should thus be interpreted with caution.

	Method	Average	Czech R.	Germany	France	U.K.	Poland
Cases	Inv. score w. median	1.003	1.019	1.156	0.991	0.891	1.177
	QRA	1.265	1.237	1.089	1.179	1.481	1.228
	QRA - adj. weights	1.168	1.163	1.009	1.121	1.281	1.288
Deaths	Inv. score w. median	1.078	1.054	1.050	0.840	1.062	1.304
	QRA	1.359	1.119	1.235	1.065	1.189	1.897
	QRA - adj. weights	1.209	0.944	1.091	1.169	1.036	1.561

Table A2: Average relative scores of ensemble forecasts that were composed by calculating weights at the level of the dominant modeling strategies within the European Hub (mechanistic, semi-mechanistic, statistical), compared to an equally weighted median based on all available models (including expert-judgment and agent-based models). The inverse score weighted median ("Inv. score w. median") calculated weights based on the average score each group of models obtained during the last four weeks, via an exponential weighting scheme that gave higher weight to recently obtained scores - the ensemble was built via a weighted median of all component forecasts (calculated weights were equally distributed between models that belonged to a group). The QRA method calculated weights for three ensembles that were built from the model groups and subsequently combined these via a weighted mean. The adjusted QRA method ("QRA - adj. weights") adjusted weights of the QRA method to account for the differing number of models within each group.

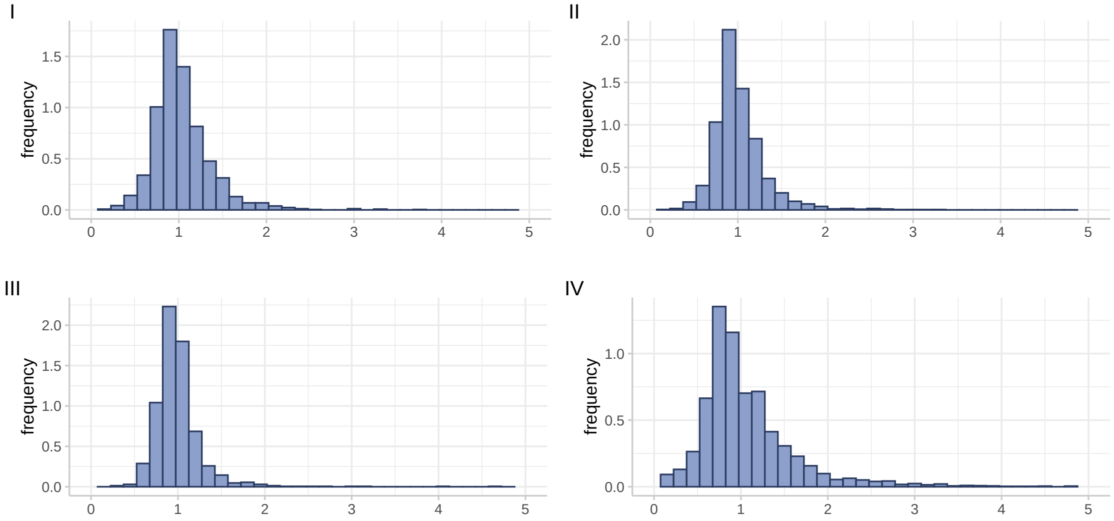


Figure A1: Exemplary conditional distributions of relative scores that result after adding a model to a base ensemble with four member models. Base ensembles were recombined from the set of available forecast models in the data at each forecast date, separately for the different target series and locations. conditions: type of added model, horizon, location, target series, period

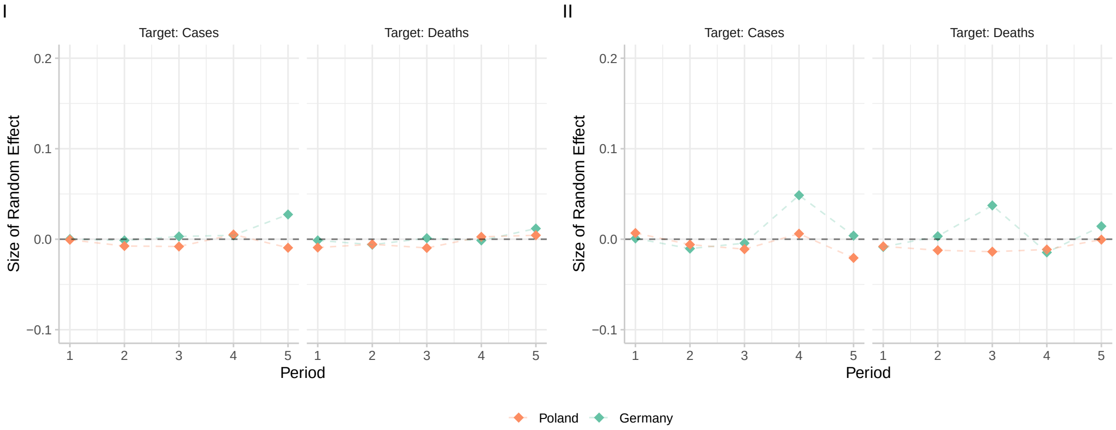


Figure A2: Fitted random effects for the mean parameter from a generalized regression with gamma distributed response. The response variable is the relative score (WIS) that results after adding a model to a base ensemble with eight member models, using the quantile-wise median (I) or mean (II) as the aggregation function. The base ensembles were recombined from the set of available forecast models in the data at each forecast date, separately for the different target series and locations.

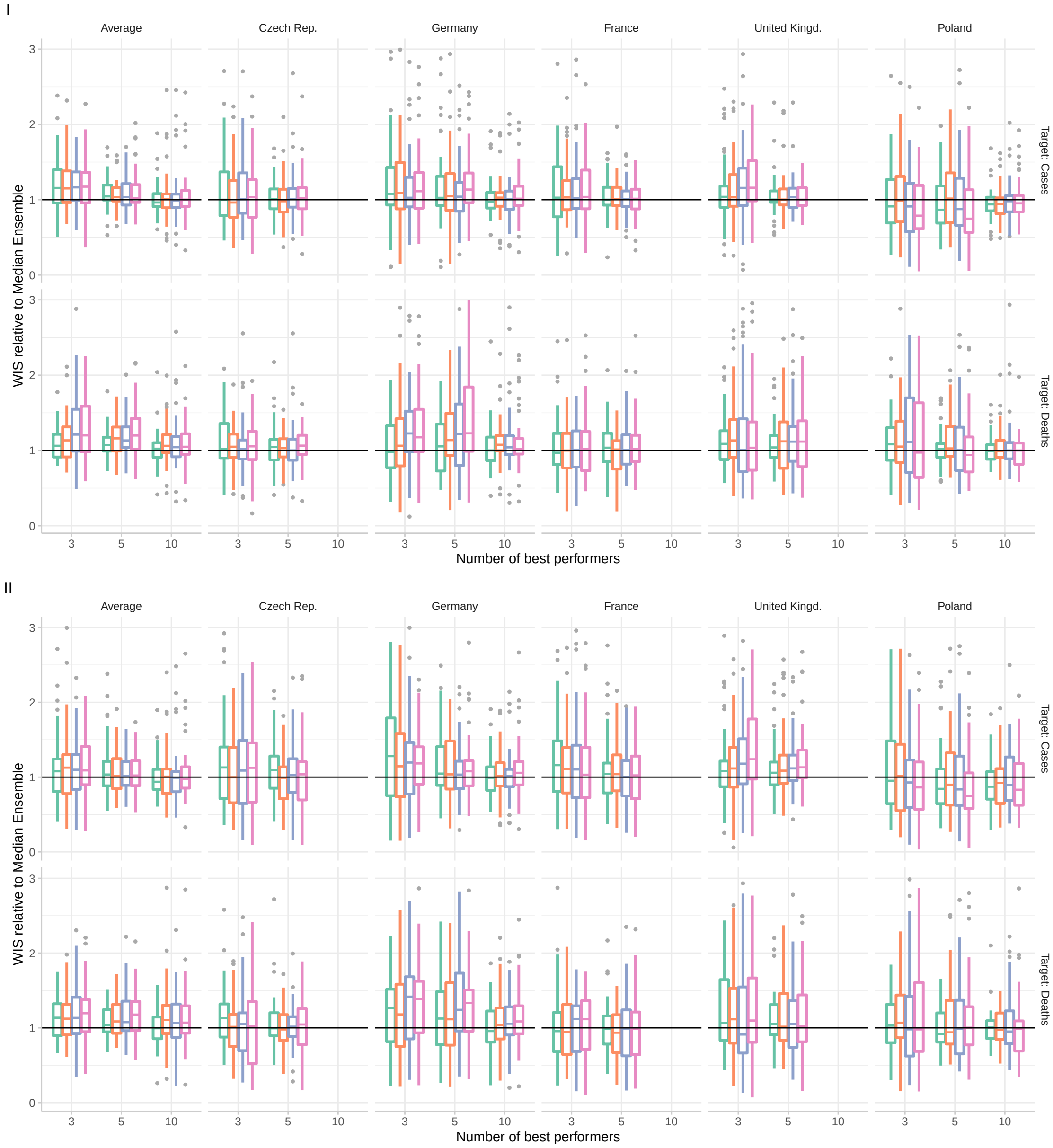


Figure A3: WIS spread of the selection ensemble methods (based on recent component forecast performance) relative to the benchmark (equally weighted median ensemble including all models), by location, target series, forecast horizon and number of best recent performers (k) included. The four panels show the resulting spreads when subsequently applying different ensemble methods to the selected component forecasts. (I): equally weighted median. (II): equally weighted mean. For all methods, selection of component forecasters for a given forecast origin is based on performance in the past four weeks. Outliers with relative scores over 3 were excluded from the plots for legibility.

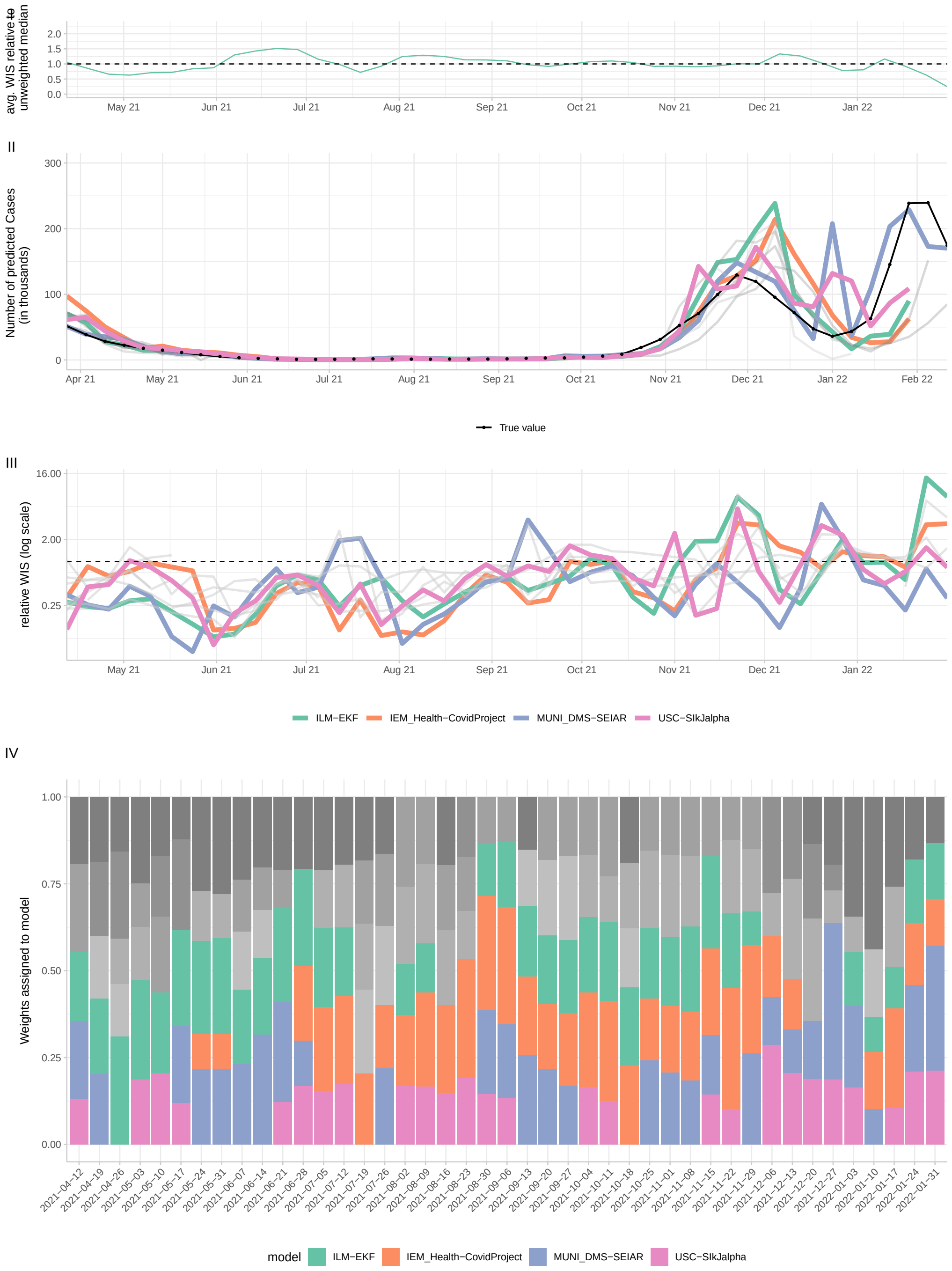


Figure A4: Performance of weekly cases forecasts for the Czech Republic, for the weighted median selection ensemble based on the five component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the four models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Furthermore, note that the y-axis of panel (I) was deliberately stretched for compatibility with panel (I) in Figure 16. Plot inspired by Ray et al. (2022).

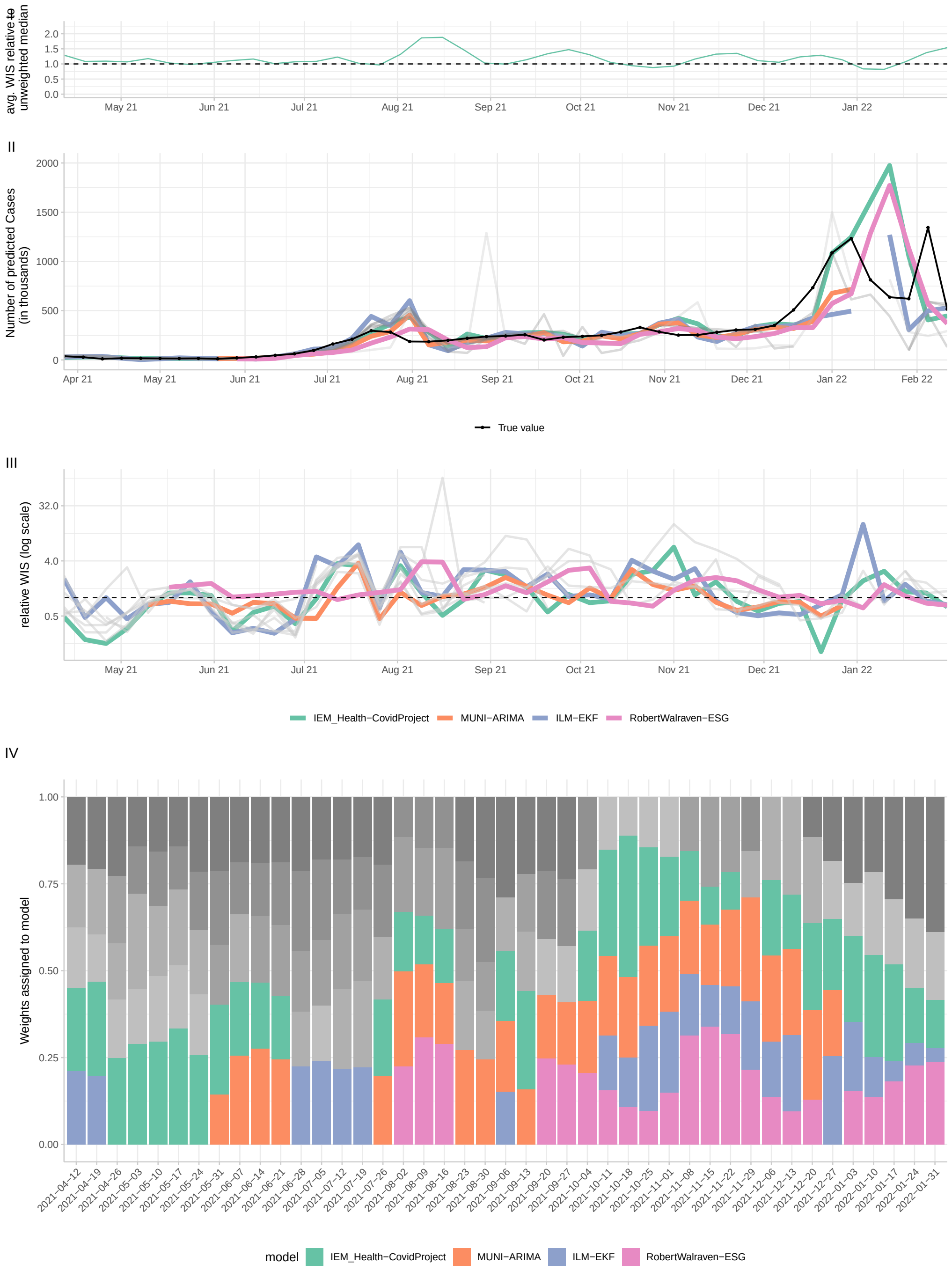


Figure A5: Performance of weekly cases forecasts for the United Kingdom, for the weighted median selection ensemble based on the five component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the four models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Furthermore, note that the y-axis of panel (I) was deliberately stretched for compatibility with panel (I) in Figure 16. Plot inspired by Ray et al. (2022).

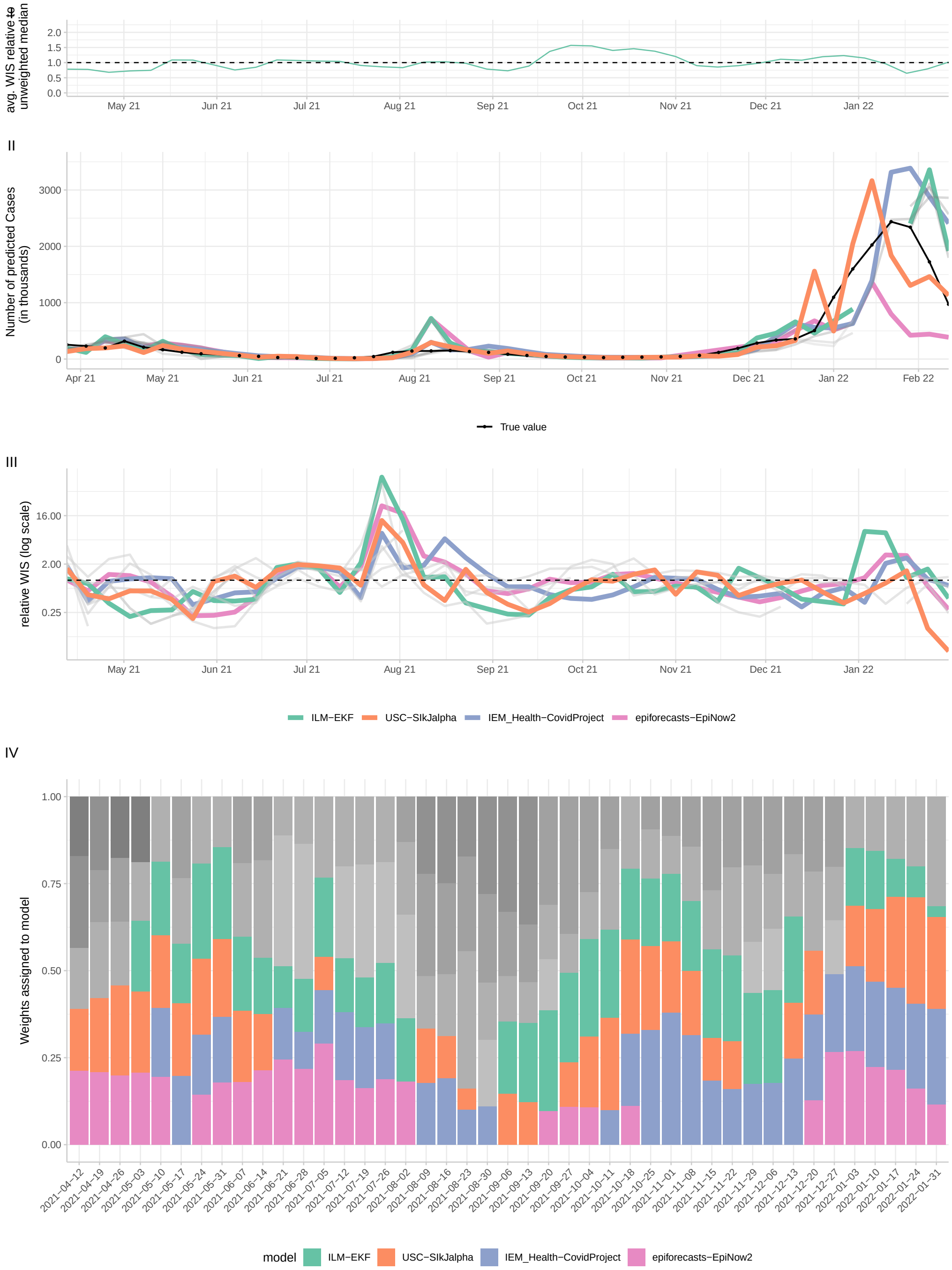


Figure A6: Performance of weekly cases forecasts for France, for the weighted median selection ensemble based on the four component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the four models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Furthermore, note that the y-axis of panel (I) was deliberately stretched for compatibility with panel (I) in Figure 16. Plot inspired by Ray et al. (2022).

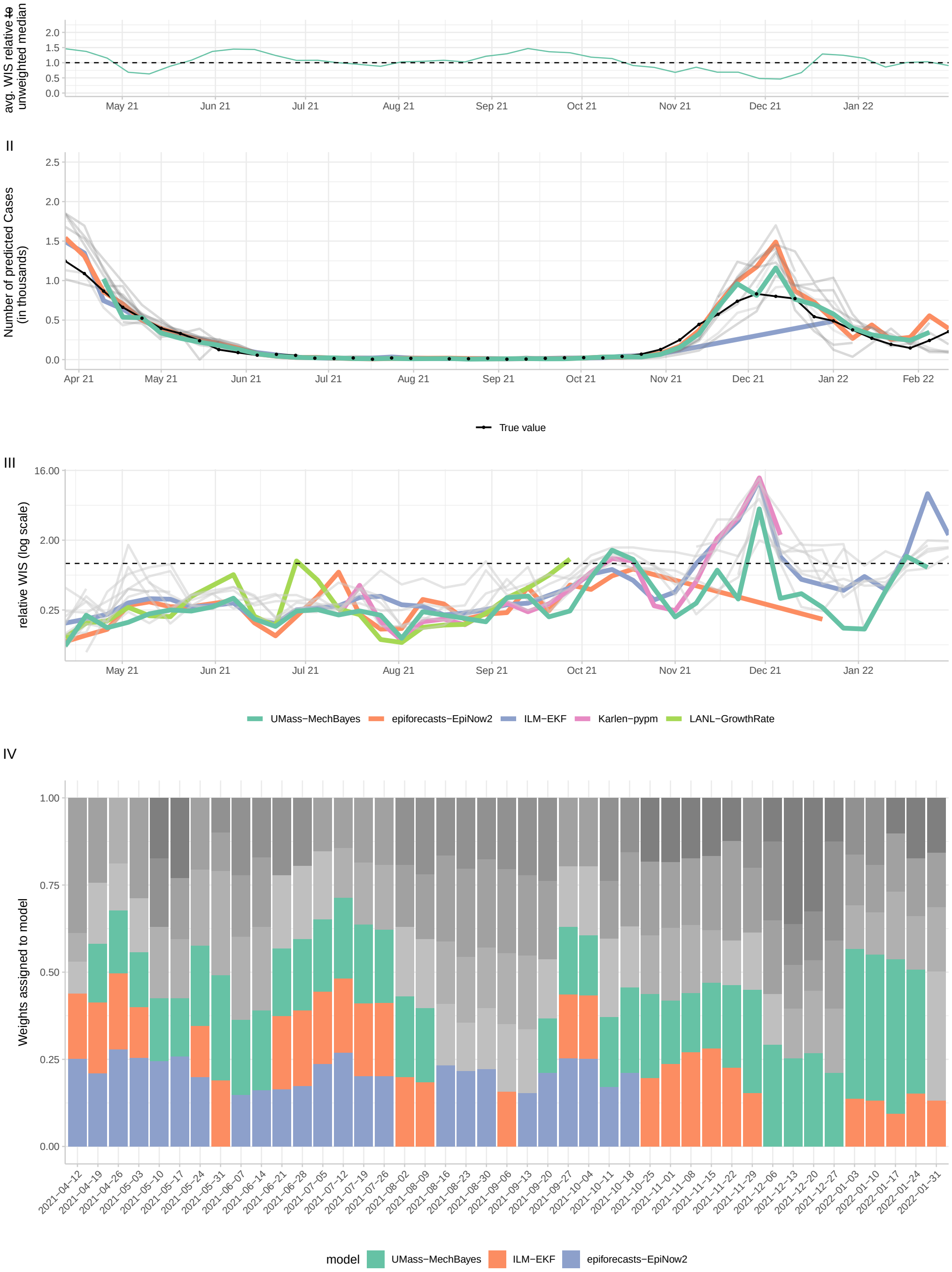


Figure A7: Performance of weekly death forecasts for the Czech Republic, for the weighted median selection ensemble based on the three component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the four models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Furthermore, note that the y-axis of panel (I) was deliberately stretched for compatibility with panel (I) in Figure 16. Plot inspired by Ray et al. (2022).

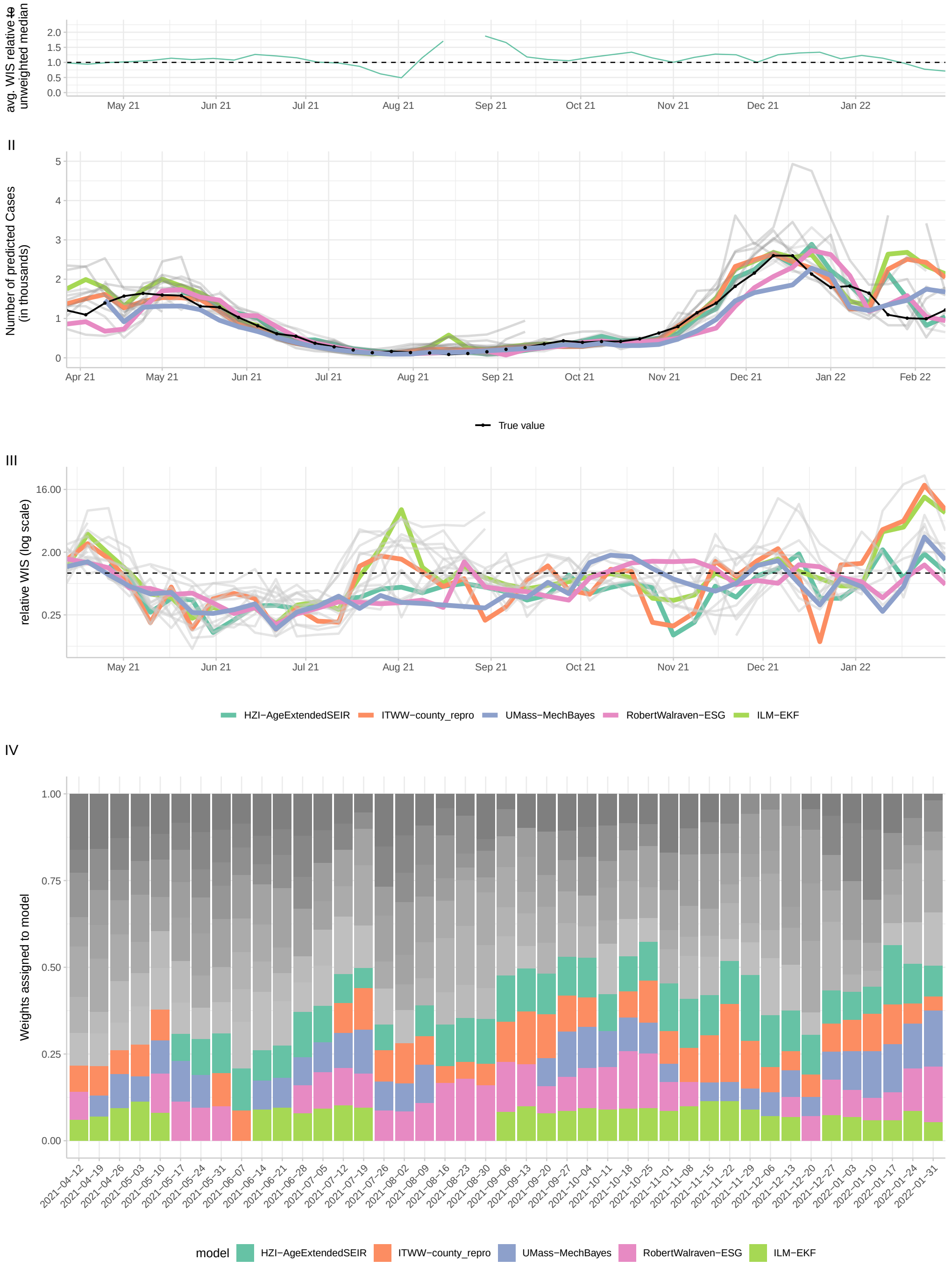


Figure A8: Performance of weekly death forecasts for Germany, for the weighted median selection ensemble based on the ten component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the five models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Furthermore, note that the y-axis of panel (I) was deliberately stretched for compatibility with panel (I) in Figure 16. Plot inspired by Ray et al. (2022).

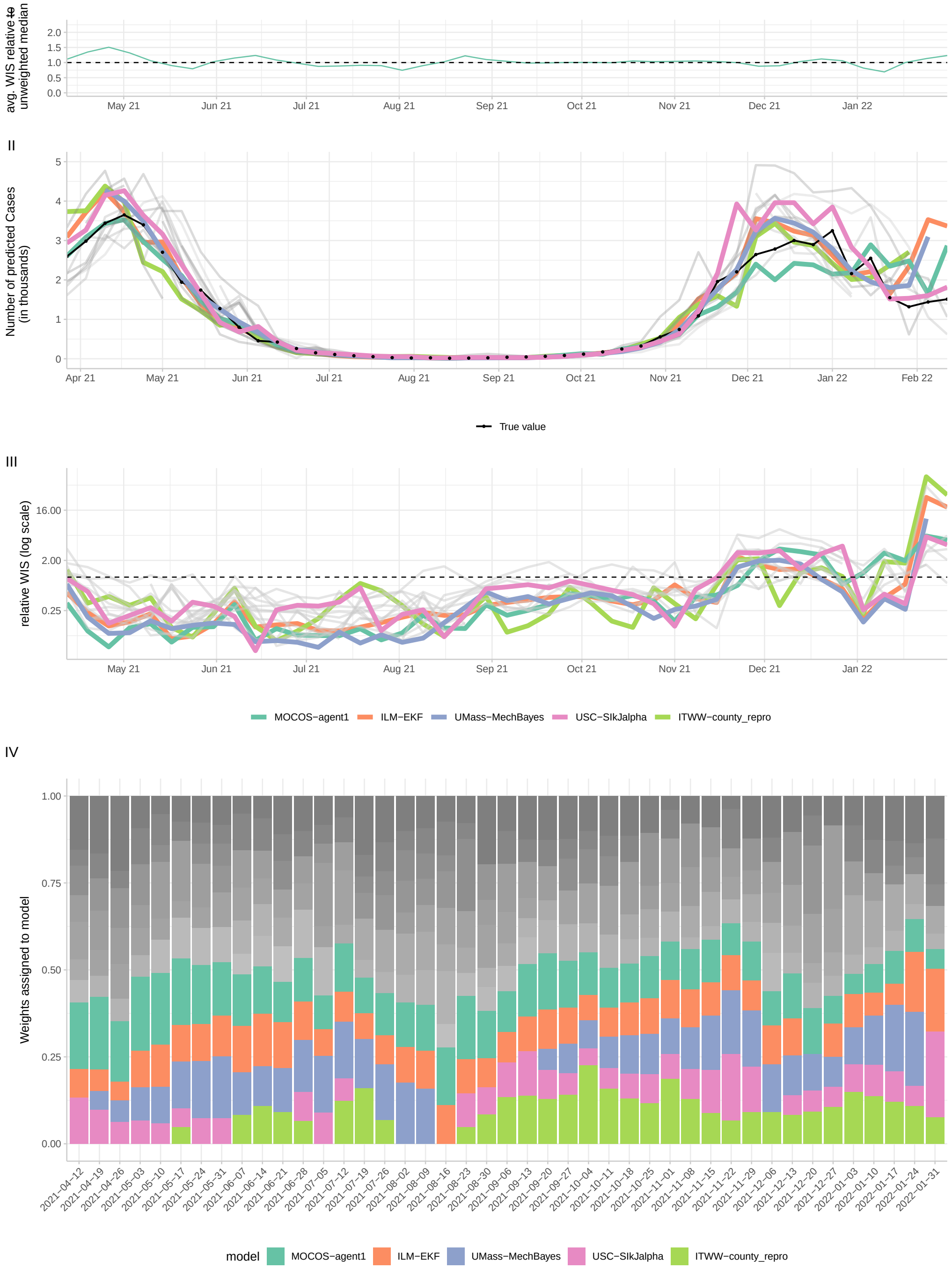


Figure A9: Performance of weekly death forecasts for Poland, for the weighted median selection ensemble based on the ten component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the five models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Furthermore, note that the y-axis of panel (I) was deliberately stretched for compatibility with panel (I) in Figure 16. Plot inspired by Ray et al. (2022).

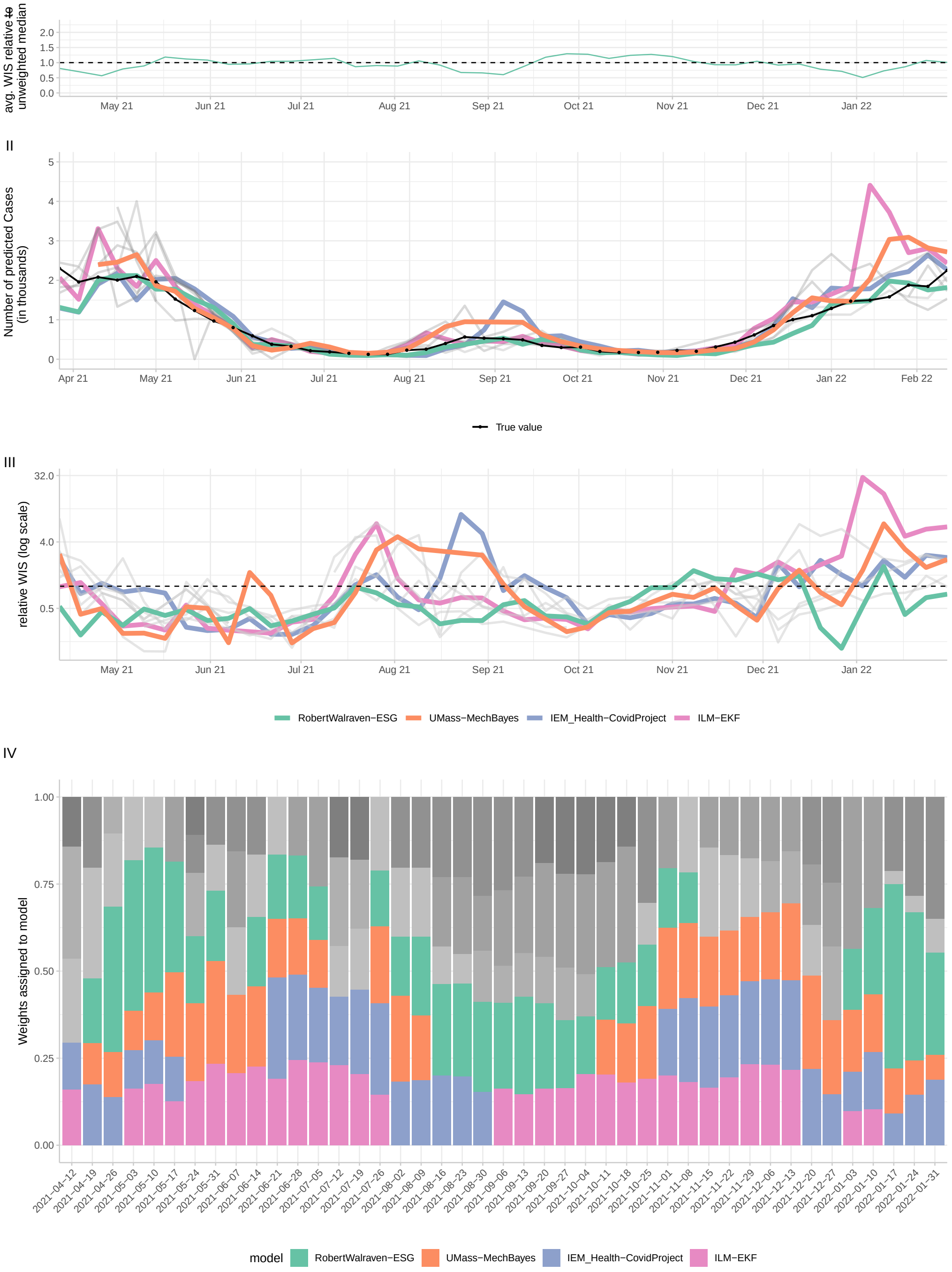


Figure A10: Performance of weekly death forecasts for France, for the weighted median selection ensemble based on the five component forecast with best performance, and its component forecasts. Component forecasts are selected and weighted based on their performance in the most recent four weeks. Across all plots, the four models chosen most often for the selection ensemble are highlighted - all other models are generally shown in gray. Panel (I) shows the relative performance of the weighted median selection ensemble based on recent component forecast performance, relative to the benchmark (unweighted median ensemble including all models). Panel (II) shows two week ahead predictions of all component forecasts, as well as the observed values of the series. Panel (III) shows performance of component forecasts over time, in terms of relative WIS. Panel (IV) shows the weekly distribution of weights given to component forecasts. Note that as forecasts are scored by the forecast date, performance measure series lead ahead of the observed incidence series. Furthermore, note that the y-axis of panel (I) was deliberately stretched for compatibility with panel (I) in Figure 16. Plot inspired by Ray et al. (2022).

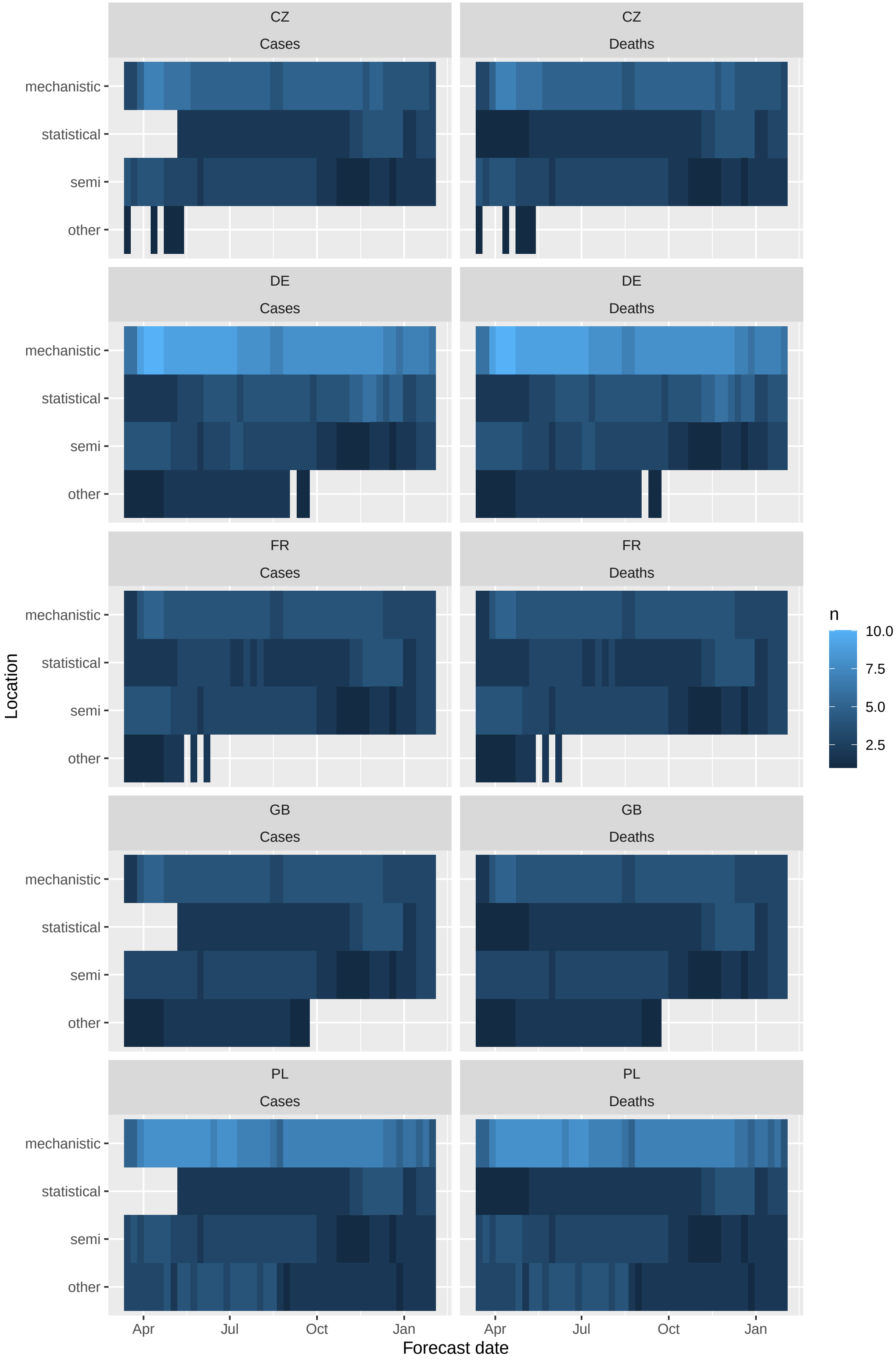


Figure A11: Availability of models for the different modeling strategies for the five location considered from the European COVID-19 Forecast Hub types, by location and target series.

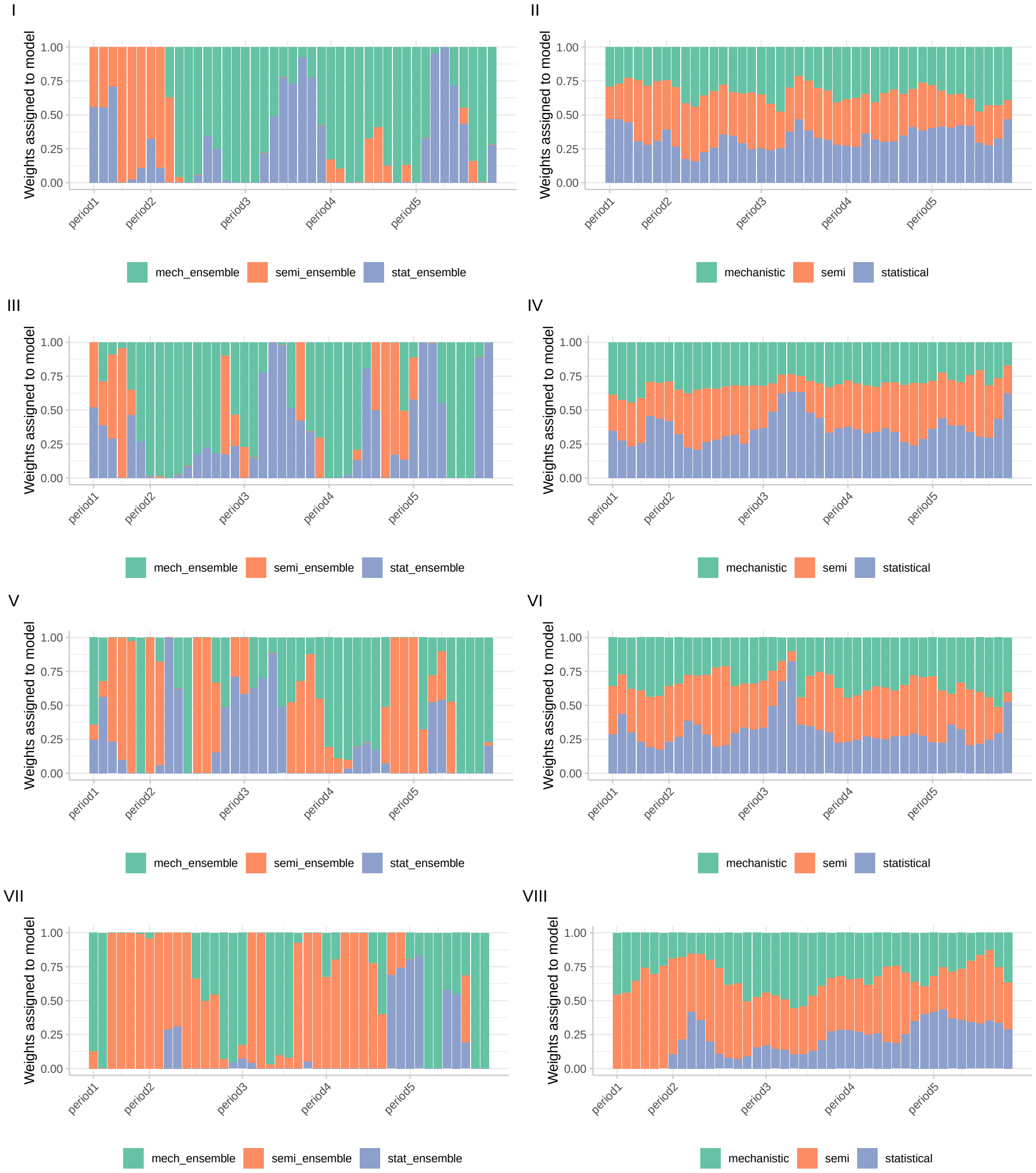


Figure A12: Weights of QRA ensemble (left column) and exponentially smoothed inverse score weight ensemble (right column), for the series: Row 1 (I, II): U.K Deaths, Row 2 (III, IV): Germany Deaths, Row 3 (V, VI): France Cases, Row 4 (VII, VIII): Poland Cases

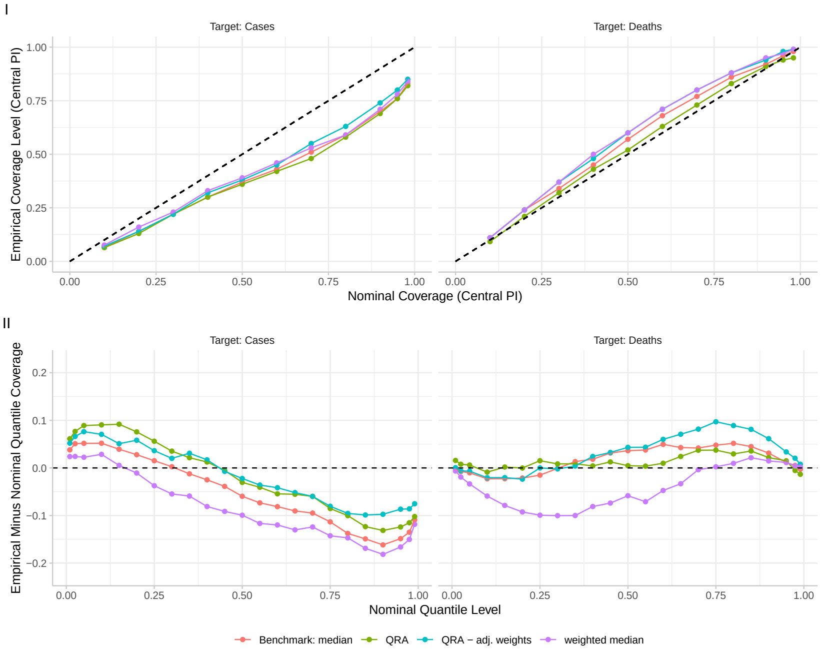
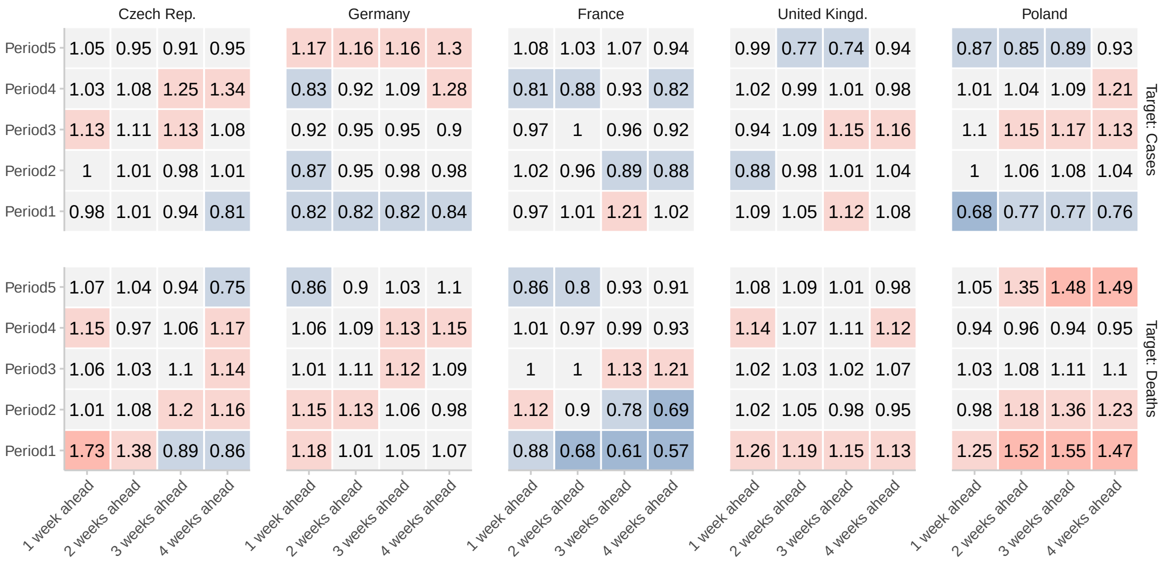


Figure A13: Assessment of probabilistic calibration of methods based on weighting at the level of recent modeling strategy performance, compared to the benchmark (equally weighted median ensemble including all models). The upper panel (I) shows central interval coverage of the methods, that is, the proportion of observations falling into the predictive distributions' central prediction intervals. The lower panel (II) shows the difference between the empirical and nominal coverage of the predictive quantiles. Negative (positive) values indicate that less (more) observations fell below a predictive distribution's respective quantile than required. For both panels, the black dashed line corresponds to optimal calibration.



Target: Cases

	Czech Rep.				Germany				France				United Kingd.				Poland			
	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead
Period5	1.05	0.95	0.91	0.95	1.17	1.16	1.16	1.3	1.08	1.03	1.07	0.94	0.99	0.77	0.74	0.94	0.87	0.85	0.89	0.93
Period4	1.03	1.08	1.25	1.34	0.83	0.92	1.09	1.28	0.81	0.88	0.93	0.82	1.02	0.99	1.01	0.98	1.01	1.04	1.09	1.21
Period3	1.13	1.11	1.13	1.08	0.92	0.95	0.95	0.9	0.97	1	0.96	0.92	0.94	1.09	1.15	1.16	1.1	1.15	1.17	1.13
Period2	1	1.01	0.98	1.01	0.87	0.95	0.98	0.98	1.02	0.96	0.89	0.88	0.88	0.98	1.01	1.04	1	1.06	1.08	1.04
Period1	0.98	1.01	0.94	0.81	0.82	0.82	0.82	0.84	0.97	1.01	1.21	1.02	1.09	1.05	1.12	1.08	0.68	0.77	0.77	0.76

Target: Deaths

	Czech Rep.				Germany				France				United Kingd.				Poland			
	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead	1 week ahead	2 weeks ahead	3 weeks ahead	4 weeks ahead
Period5	1.07	1.04	0.94	0.75	0.86	0.9	1.03	1.1	0.86	0.8	0.93	0.91	1.08	1.09	1.01	0.98	1.05	1.35	1.48	1.49
Period4	1.15	0.97	1.06	1.17	1.06	1.09	1.13	1.15	1.01	0.97	0.99	0.93	1.14	1.07	1.11	1.12	0.94	0.96	0.94	0.95
Period3	1.06	1.03	1.1	1.14	1.01	1.11	1.12	1.09	1	1	1.13	1.21	1.02	1.03	1.02	1.07	1.03	1.08	1.11	1.1
Period2	1.01	1.08	1.2	1.16	1.15	1.13	1.06	0.98	1.12	0.9	0.78	0.69	1.02	1.05	0.98	0.95	0.98	1.18	1.36	1.23
Period1	1.73	1.38	0.89	0.86	1.18	1.01	1.05	1.07	0.88	0.68	0.61	0.57	1.26	1.19	1.15	1.13	1.25	1.52	1.55	1.47

Figure A14: Average WIS of the weighted median ensemble (weights calculated by inverse score weighting at the level of modeling strategies), scaled by the WIS of an equally weighted median ensemble ("benchmark" model) based on the same set of component forecasts. Scores are calculated by location, target series, period and horizon. Relative scores above 1 correspond to the weighted median performing worse than the benchmark, while values below 1 correspond to it performing better than the benchmark. The code to produce this plot was adapted from the `scoringutils` package (Bosse et al., 2022b).