Short-term forecasting of COVID-19 in Germany and Poland during the second wave – a preregistered study

J. Bracher[1,2,*], D. Wolffram[1,2], J. Deuschel[1], K. Görgen[1], J.L. Ketterer[1], A. Ullrich[3], S. Abbott[4], M.V. Barbarossa[5], D. Bertsimas[6], S. Bhatia[7], M. Bodych[8], N.I. Bosse[4], J.P. Burgard[9], J. Fuhrmann[5,10], S. Funk[4], K. Gogolewski[11], Q. Gu[12], S. Heyder[13], T. Hotz[13], Y. Kheifetz[14], H. Kirsten[14], T. Krueger[8], E. Krymova[15], M.L. Li[16], J.H. Meinke[10], K. Niedzielewski[17], T. Ożański[8], F. Rakowski[17], M. Scholz[14], S. Soni[6], A. Srivastava[18], J. Zieliński[17], D. Zou[12], T. Gneiting[2,19], M. Schienle[1,*]

December 24, 2020

Abstract

We report insights from ten weeks of collaborative COVID-19 forecasting for Germany and Poland (12 October – 19 December 2020). The study period covers the onset of the second wave in both countries, with tightening non-pharmaceutical interventions (NPIs) and subsequently a decay (Poland) or plateau and renewed increase (Germany) in reported cases. Thirteen independent teams provided probabilistic real-time forecasts of COVID-19 cases and deaths. These were reported for lead times of one to four weeks, with evaluation focused on one- and two-week horizons, which are less affected by changing NPIs. Heterogeneity between forecasts was considerable both in terms of point predictions and forecast spread. Ensemble forecasts showed good relative performance, in particular in terms of coverage, but did not clearly dominate single-model predictions. The study was preregistered and will be followed up in future phases of the pandemic.

This is a preprint. It has not yet undergone peer review.

* Correspondence to: Johannes Bracher, (johannes.bracher@kit.edu), Melanie Schienle (melanie.schienle@kit.edu)

[1]Chair of Statistics and Econometrics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2]Computational Statistics Group, Heidelberg Institute for Theoretical Studies (HITS), Germany
[3]Robert Koch Institute (RKI), Berlin, Germany
[4]London School of Hygiene and Tropical Medicine, London, UK
[5]Frankfurt Institute for Advanced Studies, Frankfurt, Germany
[6]Sloan School of Management, Massachusetts Institute of Technology, USA
[7]MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), Imperial College London, London, UK
[8]Wroclaw University of Science and Technology, Poland
[9]Economic and Social Statistics Department, University of Trier, Germany
[10]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany
[11]Institute of Informatics, University of Warsaw, Warsaw, Poland
[12]Department of Computer Science, University of California, Los Angeles, USA
[13]Institute of Mathematics, Technische Universität Ilmenau, Germany
[14]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany
[15]Swiss Data Science Center, ETH Zurich and EPFL, Lausanne, Switzerland
[16]Operations Research Center, Massachusetts Institute of Technology, USA
[17]Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, Warsaw, Poland
[18]Ming Hsieh Department of Computer and Electrical Engineering, University of Southern California, Los Angeles, USA
[19]Institute for Stochastics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

1 Introduction

Forecasting is one of the key purposes of epidemic modelling, and despite being related to the understanding of underlying mechanisms, it is a conceptually distinct task (Keeling and Rohani, 2008). Accurate disease forecasts can improve situational awareness of decision makers and facilitate tasks such as resource allocation or planning of vaccine trials (Dean et al., 2020). During the COVID-19 pandemic, there has been a major surge in research activity on epidemic forecasting with a plethora of approaches being pursued. Contributions vary greatly in terms of purpose, forecast targets, methods, and evaluation criteria. An important distinction is between *longer-term scenario* or *what-if projections* and *short-term forecasts* (Reich and Rivers, 2020). The former attempt to discern the consequences of hypothetical scenarios and typically cannot be evaluated directly using subsequently observed data. The latter, which are the focus of this work, quantitatively describe expectations and uncertainties in the short run. They refer to quantities expected to be largely unaffected by yet unknown changes in public health interventions. This makes them particularly suitable to assess the predictive power of computational models, a need repeatedly expressed during the pandemic (Nature Publishing Group, 2020).

In this work we present results and takeaways from a collaborative and prospective short-term COVID-19 forecasting project in Germany and Poland. The evaluation period extends from 12 October 2020 (first forecasts issued) to 19 December 2020 (last observations made), thus covering the onset of the second epidemic wave in both countries. We gathered a total of 13 modelling teams from Germany, Poland, Switzerland, the United Kingdom and the United States to generate forecasts of confirmed cases and deaths in a standardized and thus comparable manner. These are publicly available in an online repository (https://github.com/KITmetricslab/covid19-forecast-hub-de) called the *German and Polish COVID-19 Forecast Hub* and can be explored interactively in a dashboard (https://kitmetricslab.github.io/forecasthub). On 8 October 2020, we deposited a study protocol (Bracher et al., 2020b) at the registry of the Open Science Foundation (OSF), predefining the study period and procedures for a prospective forecast evaluation study. Here we report on results from this effort, addressing in particular the following questions:

- At which forecast horizons can one expect to obtain reliable forecasts for various targets?
- Are the forecasts calibrated, i.e. are they able to accurately quantify their own uncertainty?
- How good is the agreement between different forecast methods?
- Are there prediction approaches which prove to be particularly reliable?
- Can combined ensemble forecasts lead to improved performance?

The study period is marked by overall strong virus circulation and changes in intervention measures and testing strategies. This makes for a situation in which reliable short-term predictions are both particularly useful and particularly challenging to produce. Conclusions from ten weeks of real-time forecasting are necessarily preliminary, but we hope to contribute to an ongoing exchange on best practices in the field. Our study will be followed up until at least March 2021 and may be extended beyond.

The project follows several principles which we consider key for a rigorous assessment of forecasting methods. Firstly, forecasts are made in real time, as retrospective forecasting often leads to overly optimistic conclusions about performance. Real-time forecasting poses many specific challenges (Desai et al., 2019), including noisy or delayed data, incomplete knowledge on testing and interventions as well as time pressure. Even if these are mimicked in retrospective studies, some benefit of hindsight remains. Secondly, in a pandemic situation with presumably low predictability we consider it of central importance to explicitly quantify forecast uncertainty. Forecasts should thus be probabilistic rather than limited to point forecasts (Held et al., 2017; Funk et al., 2019). Lastly, forecast studies are most informative if they involve statistically sound comparisons between multiple independently run forecast methods (Viboud and Vespignani, 2019). We therefore aimed for a body of standardized, comparable and uniformly formatted short-term forecasts. Such collaborative efforts have led to important advances in short-term disease forecasting prior to the pandemic (Viboud et al., 2018; Del Valle et al., 2018; Johansson et al., 2019; Reich et al., 2019a). Notably, they have provided evidence that ensemble forecasts combining various independent predictions can lead

to improved performance, similar to what has been observed in weather prediction (Gneiting and Raftery, 2005).

The German and Polish Forecast Hub project also aims to provide a platform for exchange between research teams from both countries and beyond. To this end, regular video conferences with presentations and discussions on forecast methodologies were organized. Moreover, the Forecast Hub Team provided feedback on performance in order to facilitate model revisions and forecast improvement.

The German and Polish COVID-19 Forecast Hub is run in close exchange with the US COVID-19 Forecast Hub (Ray et al. 2020; COVID-19 Forecast Hub Team 2020) and aims for compatibility with the short-term forecasts assembled there. Consequently, many formal aspects presented in Section 2 are shared between the two projects. However, we faced a number of distinct challenges, including rapid changes in non-pharmaceutical interventions, the use of different truth data sources by different teams and a smaller number of contributing teams. Close links moreover exist to a similar effort in the United Kingdom (Funk et al., 2020). Other conceptually related works on short-term forecasting or baseline projections include those by the Austrian COVID-19 Forecast Consortium (Bicher et al., 2020) and the European Centre for Disease Prevention and Control (ECDC; 2020a; 2020c). In a German context, various nowcasting efforts exist, see e.g. Günther et al. (2020).

2 Formal setting

We start by laying out the formal framework of the presented collaborative forecasting study. Unless stated differently, the principles correspond to those specified in the study protocol (Bracher et al., 2020b).

2.1 Submission system and rhythm

All submissions were collected in a standardized format in a public repository to which teams could submit (https://github.com/KITmetricslab/covid19-forecast-hub-de). For teams running their own repositories, the Forecast Hub Team put in place software scripts to re-format forecasts and transfer them into the Hub repository. Participating teams were asked to update their forecasts on a weekly basis using data up to Monday. Submission was possible until Tuesday 3 pm Berlin/Warsaw time. Delayed submission of forecasts was possible until Wednesday, with exceptional further extensions possible in case of technical issues. Delays of submissions were documented (Supplementary Table 4).

2.2 Forecast targets and format

We focus on short-term forecasting of confirmed cases and deaths from COVID-19 in Germany and Poland one and two weeks ahead. Here, weeks refer to Morbidity and Mortality Weekly Report (MMWR) weeks which start on Sunday and end on Saturday, meaning that one-week-ahead forecasts were actually five days ahead, two-week ahead forecasts were twelve days ahead, etc. All targets were defined by the date of reporting to the national authorities (rather than e.g. symptom onset date). This means that modellers have to take reporting delays into account, but has the advantage that data points are usually not revised over the following days and weeks. All targets were addressed both on cumulative and weekly incident scales. Forecasts could refer to both data from the European Centre for Disease Prevention and Control (ECDC; 2020b) and Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE; Dong et al. 2020). In this article we focus on the preregistered period of 12 October 2020 to 19 December 2020. Figure 1 shows the targets on an incidence scale for the two countries, with the study period highlighted. We also indicate the timing of changes in interventions and reporting procedures which were considered of importance for short-term forecasting.

Note that on 14 December 2020, the ECDC data set on COVID-19 cases and deaths in daily resolution was discontinued. For the last weekly data point we therefore used data streams from Robert Koch Institute and the Polish Ministry of Health which we had previously used to obtain regional data and which up to this time had been in agreement with the ECDC data.

Most forecasters also produced and submitted three- and four-week-ahead forecasts (which were specified as targets in the study protocol). These horizons, also used in the US COVID-19 Forecast Hub (Ray et al.,

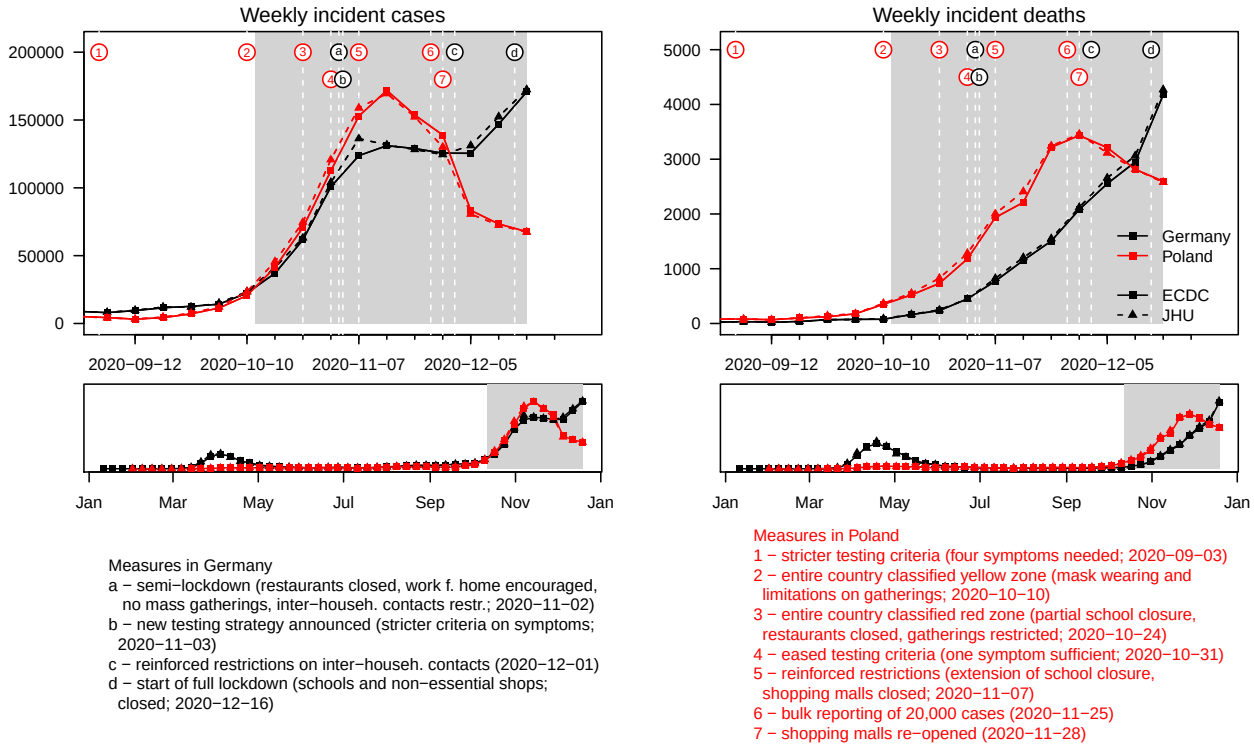

Figure 1: Weekly incident confirmed cases and deaths from COVID-19 in Germany and Poland according to data sets from ECDC and JHU. The study period covered in this paper is highlighted in grey. Important changes in interventions and testing are marked. Sources containing details on the listed interventions are provided in Supplementary Section C.

2020), were originally defined for deaths. Due to their lagged nature, these were considered predictable independently of future policy or behavioural changes up to four weeks ahead; see UK Scientific Pandemic Influenza Group on Modelling (2020) for a similar argument. During the summer months, when incidence was low and intervention measures largely constant, the same horizons were introduced for cases. As the epidemic situation and intervention measures became more dynamic in autumn, it became clear that case forecasts further than two weeks (twelve days) ahead were too dependent on yet unknown interventions and the consequent changes in transmission rates. It was therefore decided to restrict the default view in the online dashboard to one- and two-week-ahead forecasts only. At the same time we continued to collect three- and four-week-ahead outputs. Most models (with the exception of `epiforecasts-EpiExpert`, `COVIDAnalytics-Delphi` and in some exceptional cases `MOCOS-agent1`) do not anticipate policy changes, so that their outputs can be seen as "baseline projections", i.e. projections for a scenario with constant interventions. In accordance with the study protocol we also report on three- and four-week-ahead predictions, but these results have been moved to the Supplementary Material.

We emphasize the importance of quantifying the uncertainty associated with forecasts. Teams were therefore asked to report a total of 23 predictive quantiles (1%, 2.5%, 5%, 10%, ..., 90%, 95%, 97.5%, 99%) in addition to their point forecasts. This motivates considering both forecasts of cumulative and incident quantities, as predictive quantiles for these generally cannot be translated from one scale to the other. Not all teams provided such probabilistic forecasts, though, and we also accepted pure point forecasts.

2.3 Evaluation measures

The submitted quantiles of a predictive distribution F define 11 central prediction intervals with nominal coverage level $1-\alpha$ where $\alpha = 0.02, 0.05, 0.10, 0.20, \ldots, 0.90$. Each of these can be evaluated using the *interval score* (Gneiting and Raftery, 2007):

$$\mathrm{IS}_\alpha(F, y) = (u - l) \; + \; \frac{2}{\alpha} \times (l - y) \times \mathbf{1}(y < l) \; + \; \frac{2}{\alpha} \times (y - u) \times \mathbf{1}(y > u).$$

Here u and l are the lower and upper ends of the respective interval, $\mathbf{1}$ is the indicator function and y is the eventually observed value. The three summands can be interpreted as a measure of sharpness and penalties for under- and overprediction, respectively. The primary evaluation measure used in this study is the *weighted interval score* (WIS; Bracher et al. 2020a), which combines the absolute error (AE) of the predictive median m and the interval scores achieved for the eleven nominal levels. The WIS is a well-known quantile-based approximation of the continuous ranked probability score (CRPS; Gneiting and Raftery 2007) and, in the case of our 11 intervals, defined as

$$\mathrm{WIS}(F, y) = \frac{1}{12} \times \left(|y - m| \; + \; \sum_{k=1}^{11} \left(\frac{\alpha_k}{2} \times \mathrm{IS}_{\alpha_k}(F, y) \right) \right),$$

where $\alpha_1 = 0.02, \alpha_2 = 0.05, \alpha_3 = 0.10, \alpha_4 = 0.20, \ldots, \alpha_{11} = 0.90$. The score reflects the distance between the predictive distribution F and the eventually observed outcome y, and thus is negatively oriented, meaning that smaller values are better. As secondary measures of forecast performance we considered the absolute error of point forecasts and the empirical coverage of 50% and 95% prediction intervals. In this context we note that WIS and AE are equivalent for deterministic forecasts (i.e. forecasts concentrating all probability mass on a single value). This enables a principled comparison between probabilistic and deterministic forecasts, both of which appear in the present study.

In the evaluation we needed to account for the fact that forecasts can refer to either the ECDC or JHU data sets. We performed all forecast evaluations once using ECDC data and once using JHU data, with ECDC being our prespecified primary data source. For cumulative targets we shifted forecasts which refer to the other truth data source additively by the last observed difference. This is a pragmatic strategy to align forecasts with the last state of the respective time series.

Another difficulty in comparative forecast evaluation lies in the handling of missing forecasts. For this case (which indeed occurred for several teams) we prespecified that the missing score would be imputed with the worst (i.e. largest) score obtained by any other forecaster. In the respective summary tables any such instances are marked. All values reported are mean scores over the evaluation period, though if more than a third of the forecasts were missing we refrain from reporting.

3 Forecasting methods

3.1 Baseline forecasts

In order to put evaluation results into perspective we use three simple reference models. Note that only the first was prespecified. The two others were added later as the need for comparisons to simple, but not completely naïve, approaches was recognized. More detailed descriptions are provided in Supplementary Section B.

`KIT-baseline`: A naïve last-observation carried-forward approach (on the incidence scale) with identical variability for all forecast horizons (estimated from the last five observations). This is very similar to the *null model* used by Funk et al. (2020).

`KIT-extrapolation_baseline`: A multiplicative extrapolation based on the last two observations with uncertainty bands estimated from five preceding observations.

`KIT-time_series_baseline` An exponential smoothing model with multiplicative error terms and no seasonality as implemented in the `R` package `forecast` (Hyndman and Khandakar, 2008) and used for COVID-19 forecasting by Petropoulos and Makridakis (2020).

3.2 Contributed forecasts

During the evaluation period from October to December 2020, we assembled short-term predictions from a total of 14 forecast methods by 13 independent teams of researchers. Eight of these are run by teams collaborating directly with the Hub, based on models these researchers were either already running or set up specifically for the purpose of short-term forecasting. The remaining short-term forecasts were made available via dedicated online dashboards by their respective authors, often along with forecasts for other countries. With their permission, the Forecast Hub team assembled and integrated these forecasts. Table 1 provides an overview of all included models with brief descriptions and information on the handling of non-pharmaceutical interventions, testing strategies, age strata and the source used for truth data. The models span a wide range of approaches, from computationally expensive agent-based simulations to human judgement forecasts. Not all models addressed all targets and forecast horizons suggested in our project; which targets were addressed by which models can be seen from Tables 2 and 3.

3.3 Ensemble forecasts

Evidence from past forecasting efforts on various diseases (e.g., Yamana et al. 2016; Viboud et al. 2018; Reich et al. 2019b) and recent research in the context of COVID-19 (Brooks et al., 2020; Funk et al., 2020) suggests that ensemble forecasts combining several independent forecasts can lead to improved and more stable performance. We therefore assess the performance of three different forecast aggregation approaches:

`KITCOVIDhub-median_ensemble` The α-quantile of the ensemble forecast for a given quantity is given by the median of the respective α-quantiles of the member forecasts. The associated point forecast is the quantile at level $\alpha = 0.50$ of the ensemble forecast (same for other ensemble approaches).

`KITCOVIDhub-mean_ensemble` The α-quantile of the ensemble forecast for a given quantity is given by the mean of the respective α-quantiles of the member forecasts.

`KITCOVIDhub-inverse_wis_ensemble` The α-quantile of the ensemble forecast is a weighted average of the α-quantiles of the member forecasts. The weights are chosen inversely to the mean WIS value obtained by the member models over six recently evaluated forecasts (last three one-week-ahead, last two two-week-ahead, last three-week-ahead). This is done separately for incident and cumulative forecasts. The inverse-WIS ensemble is a pragmatic strategy to base weights on past performance which is feasible with a limited amount of historical forecast/observation pairs (see Zamo et al. 2020 for a similar approach).

Only models providing complete probabilistic forecasts with 23 quantiles for all four forecast horizons were included into the ensemble for a given target. It was not required that forecasts be submitted for both cumulative and incident targets, so that ensembles for incident and cumulative cases were not necessarily based on exactly the same set of models. The Forecast Hub Team reserved the right to screen and exclude member models in case of implausibilities. Decisions on inclusion were taken simultaneously for all three ensemble versions and were documented in the Forecast Hub platform. The main reasons for the exclusion of forecasts from the ensemble were forecasts in an implausible order of magnitude or forecasts with vanishingly small or excessive uncertainty. As it showed comparable performance to submitted forecasts, the `KIT-time_series_baseline` model was included in the ensemble forecasts in most weeks.

Preliminary results from the US COVID-19 Forecast Hub indicate better forecast performance of the median compared to the mean ensemble (Taylor and Taylor, 2020), and the median ensemble has served as the operational ensemble since 28 July 2020. Up to date, trained ensembles yield only limited, if any, benefits (Brooks et al., 2020). We therefore prespecified the median ensemble as our main ensemble approach. Note that in the context of influenza forecasting (Reich et al., 2019b), ensembles have been constructed by combining probability densities rather than quantiles. These approaches have somewhat different behaviour, but no general statement can be made which one yields better performance (Lichtendahl et al., 2013). As in our setting member forecasts were reported in a quantile format we resort to quantile-based methods for aggregation.

Table 1: Forecast models contributed by independent external research teams. Abbreviations: NPI: Does the forecast model explicitly account for non-pharmaceutical interventions? Test: Does the model account for changing testing strategies? Age: Is the model age-structured? DE, PL: Are forecasts issued for Germany and Poland, respectively? Truth: Which truth data source does the model use? Pr: Are forecasts probabilistic (23 quantiles)?

Category	Model	Description	NPI	Test	Age	DE	PL	Truth	Pr
Agent-based	`ICM-agentModel`	Agent-based model for stochastic simulations of air-borne disease spread. Agents are assigned to geographically distributed contexts. The model implements a travel module that moves agents between cities (Rakowski et al., 2010).	✓	✓	✓		✓	JHU	✓
	`MOCOS-agent1`	Agent based model. Continuous-time stochastic microsimulation based on census data, including contact tracing, testing and quarantine (Adamik et al., 2020). Relevant duration time distributions are based on empirical data.	✓	✓	✓		✓	JHU	✓
Compartment	`CovidAnalytics-DELPHI`[1]	Country-level modified SEIR model accounting for changing interventions and underdetection (Li et al., 2020).	✓			✓	✓	JHU	✓
	`FIAS_FZJ-Epi1Ger`	Country-level deterministic model, extension of classical SEIR approach, takes explicitly into account undetected cases and reporting delays (Barbarossa et al., 2020).				✓		ECDC	✓
	`LeipzigIMISE-SECIR`	An extension of the SECIR type implemented as input-output non-linear dynamical system. Joint fit of data on test positives, deaths, and ICU occupancy accounting for reporting delays.	✓			✓		ECDC	✓
	`MIMUW-StochSEIR`	SEIR model with extensions: introduction of the undiagnosed compartment; testing limits influencing number of diagnosed cases; stochastic perturbations of time-dependent contact rate.					✓	JHU	✓
	`UCLA-SuEIR`[2]	A variant of the SEIR model considering both untested and unreported cases (Zou et al., 2020). The model considers reopening and assumes the susceptible population will increase after the reopen.	✓	✓		✓		JHU	
	`USC-SIkJalpha`[3]	Reduces a heterogeneous rate model into multiple simple linear regression problems. True susceptible population is identified based on reported cases, whenever possible. (Srivastava et al., 2020).				✓	✓	JHU	
Growth rate/ renewal eq.	`epiforecasts-EpiNow2`	An exponential growth model that uses a time-varying R_t trajectory to forecast latent infections, then convolves these using known delays to observations. (Abbott et al., 2020). Beyond the forecast horizon R_t is assumed to be static.				✓	✓	ECDC	✓
	`Geneva-DetGrowth`[4]	Robust seasonal trend decomposition for smoothing of daily observations with further linear or multiplicative extrapolation.				✓	✓	ECDC	
	`ITWW-county_repro`	Forecasts of county level incidence based on regional reproduction numbers estimated via small area estimation.			✓	✓	✓	ECDC	✓
	`LANL-GrowthRate`[5]	Dynamic SI model for cases with growth rate parameter updated at each model run (via regression model with day-of-week effect). The deaths forecast is a fraction of the cases forecasts (fraction learned via regression and updated at each run).				✓	✓	JHU	✓
Human jugdement	`epiforecasts-EpiExpert`	A mean ensemble of predictions from experts and non-experts. Predictions are made through a web app[6] by choosing a distribution and specifying the median and width of that predictive distribution.	(✓)	(✓)	(✓)	✓	✓	ECDC	✓
Forecast ensemble	`Imperial-ensemble2`[7]	Unweighted average of four forecasts for death counts (see reference in footnote).				✓	✓	ECDC	✓

Teams marked with footnotes run their own dashboards: [1]https://www.covidanalytics.io, [2]https://covid19.uclaml.org, [3]https://scc-usc.github.io/ReCOVER-COVID-19, [4]https://renkulab.shinyapps.io/COVID-19-Epidemic-Forecasting, [5]https://covid-19.bsvgateway.org, [6]https://cmmid-lshtm.shinyapps.io/crowd-forecast, [7]https://mrc-ide.github.io/covid19-short-term-forecasts

4 Results

We start by discussing some general observations made during the evaluation period, shedding light on challenges and particularities of collaborative real-time forecasting during a pandemic. Subsequently, we provide a quantitative evaluation in terms of WIS, AE, and interval coverage. Visualizations of one- and two-week-ahead forecasts on the incidence scale are displayed in Figures 3 and 4, respectively, and will be discussed in the following subsections. Note that these figures are restricted to models submitted over (almost) the entire evaluation period and providing complete forecasts including 23 predictive quantiles. Forecasts from the remaining models are illustrated in Supplementary Section E. Forecasts at prediction horizons of three and four weeks are shown in Supplementary Section F.

4.1 Specific observations and challenges

A recurring theme during the evaluation period was pronounced variability between model forecasts. Figure 2 illustrates this aspect for point forecasts of incident cases in Germany, but it also holds for Poland and death forecasts. The left panel shows the spread of forecasts issued on 19 October 2020 and valid one to four weeks ahead. The models present very different outlooks, ranging from a return to the lower incidence of previous weeks to exponential growth. The graph also illustrates the difficulty of forecasting cases more than two weeks ahead. Several models had correctly picked up the upwards trend, but presumably a combination of the new testing regime and the semi-lockdown (marked as (a) and (b)) led to a flattening of the curve. The right panel shows forecasts from 9 November 2020, immediately following the aforementioned events. Again, the forecasts are quite heterogeneous. The week ending on Saturday 7 November had seen a slower increase in reported cases than anticipated by almost all models (see Figure 3), but there was general uncertainty about the role of saturating testing capacities and evolving testing strategies. Indeed, on 18 November it was argued in a situation report from Robert Koch Institute (RKI) that comparability of data from calendar week 46 (9–15 November) to previous weeks was limited (Robert Koch Institute, 2020). This illustrates that confirmed cases can be a moving target, and that different modelling decisions can lead to very different forecasts.

Far from all forecast models explicitly account for interventions and testing strategies (Table 1). Many forecasters instead prefer to let their models pick up trends from the data once they become apparent. This can lead to delayed adaptation to changes and explains why numerous models – including the ensemble – showed overshoot in the first half of November when cases started to plateau in Germany (visible from Figure 3 and even more pronounced in Figure 4). Interestingly, some models adapted more quickly to the flatter curve. This includes the human judgement approach `EpiExpert`, which, due to its reliance on human input and knowledge, can take information on interventions into account before they become apparent in epidemiological data, but interestingly also `Epi1Ger` and `EpiNow2` which do not account for interventions. In Poland, overshoot could be observed following the peak week in cases (ending on 15 November), with the one-week-ahead median ensemble only barely covering the next observed value. However, most models adapted quickly and were back on track in the following week.

A noteworthy difficulty for death forecasts in Germany was under-prediction in consecutive weeks in late November and December. In November, several models predicted that death numbers would stop increasing, likely as a consequence of the plateau in overall case numbers starting several weeks before. In the last week of our study (ending on 19 December) most models considerably under-estimated the increase in weekly deaths. A difficulty may have been that despite the overall plateau which was observed until early December, cases continued to increase in the oldest age groups, for which the mortality risk is highest (see Supplementary Figure 8). Models that do not take into account the age structure of cases – which includes most available models (Table 1) – may then have been led astray.

Forecasts are not only heterogeneous with respect to their point forecasts, but also the implied uncertainty. As can be seen from Figures 3 and 4, certain models issue very confident forecasts with narrow forecast intervals barely visible in the plot. Others – in particular the exponential smoothing time series model `KIT-time_series_baseline`, but also `LANL-GrowthRate` – show rather large uncertainty. For almost all forecast dates there are pairs of models with no or minimal overlap in 95% prediction intervals, another indicator of limited agreement between forecasts. As can be seen from the right column of Figures 3 and 4 as well as Tables 2 and 3, most contributed models were overconfident, i.e. their prediction intervals did not

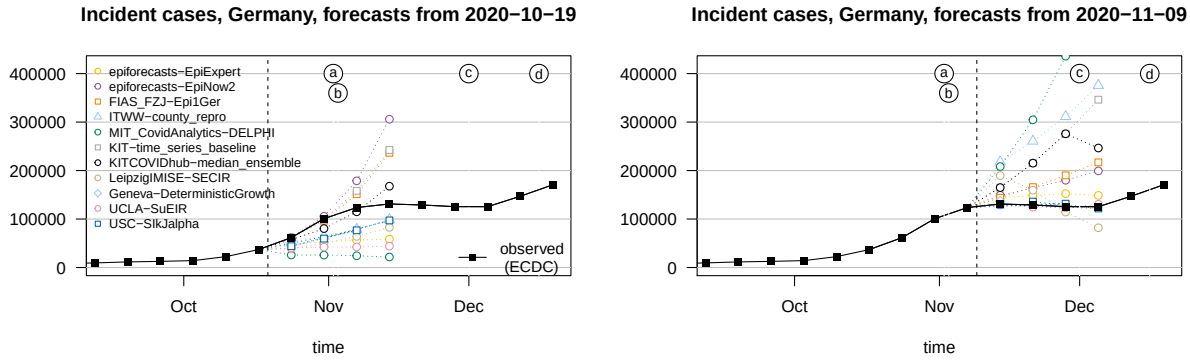

Figure 2: Illustration of heterogeneity between incident case forecasts in Germany. Left: Point forecasts issued by different models and the median ensemble on 19 October 2020. Right: Point forecasts issued on 9 November 2020. The dashed vertical line indicates the date at which forecasts were issued. Events marked by letters a – d are explained in Figure 1.

reach nominal coverage.

A major question in pandemic real-time forecasting is how closely surveillance data reflect the underlying dynamics. Like in Germany, testing criteria were repeatedly adapted in Poland. In early September they were tightened, requiring the simultaneous presence of four symptoms for the administration of a test. This was changed to less restrictive criteria in late October (presence of one characteristic symptom alone sufficient). These changes limit comparability of numbers across time. Very high test positivity rates in Poland suggest that there was substantial under-ascertainment, which is assumed to have aggravated over time. Comparisons between overall excess mortality and reported COVID deaths suggest that there is also relevant under-ascertainment of deaths, again likely changing over time (Afelt et al., 2020). These aspects make predictions challenging, and limitations of ground truth data sources are inherited by the forecasts which refer to them. A particularly striking example of this was the belated addition of 22,000 cases from previous weeks to the Polish record on 24 November 2020. We are aware that certain teams (namely, the Poland-based teams `MOCOS` and `MIMUW`) explicitly took this shift into account while others did not. This incident was not specifically accounted for in the evaluation as it was considered part of the general uncertainty affecting the prediction targets.

4.2 Findings for median, mean and inverse-WIS ensembles

Beyond comparing and evaluating short-term forecasts, we assessed the potential of forecast ensembles. Before providing a quantitative assessment in the following section, we present some general observations on the median, mean and inverse-WIS ensembles introduced in Section 3.3.

A key advantage of the median ensemble is that it is more robust to single extreme forecasts than the mean ensemble. As an example of the different behaviour in cases where one forecast differs considerably from the others we show forecasts of incident deaths in Poland from 30 November 2020 in Figure 5. The first panel shows the six member forecasts, the second the resulting median and mean ensembles. While the two ensemble forecasts are not drastically different and imply rather similar ranges, the predictive median of the latter is noticeably higher. The reason is that it is more strongly impacted by one model which predicted a resurge in deaths.

While the robustness of the median ensemble is often an advantage, we also encountered a downside of the approach. When member forecasts are rather heterogeneous, and there are low to medium numbers of members only, median ensemble forecasts are not always very well-shaped. One of the most pronounced examples we encountered is shown in the third and fourth panel of Figure 5. For the one-week-ahead forecast of incident cases in Poland from 2 November 2020, the predictive 25% quantile and median were almost identical. For the two-week-ahead median ensemble forecast, the 50% and 75% quantile were almost

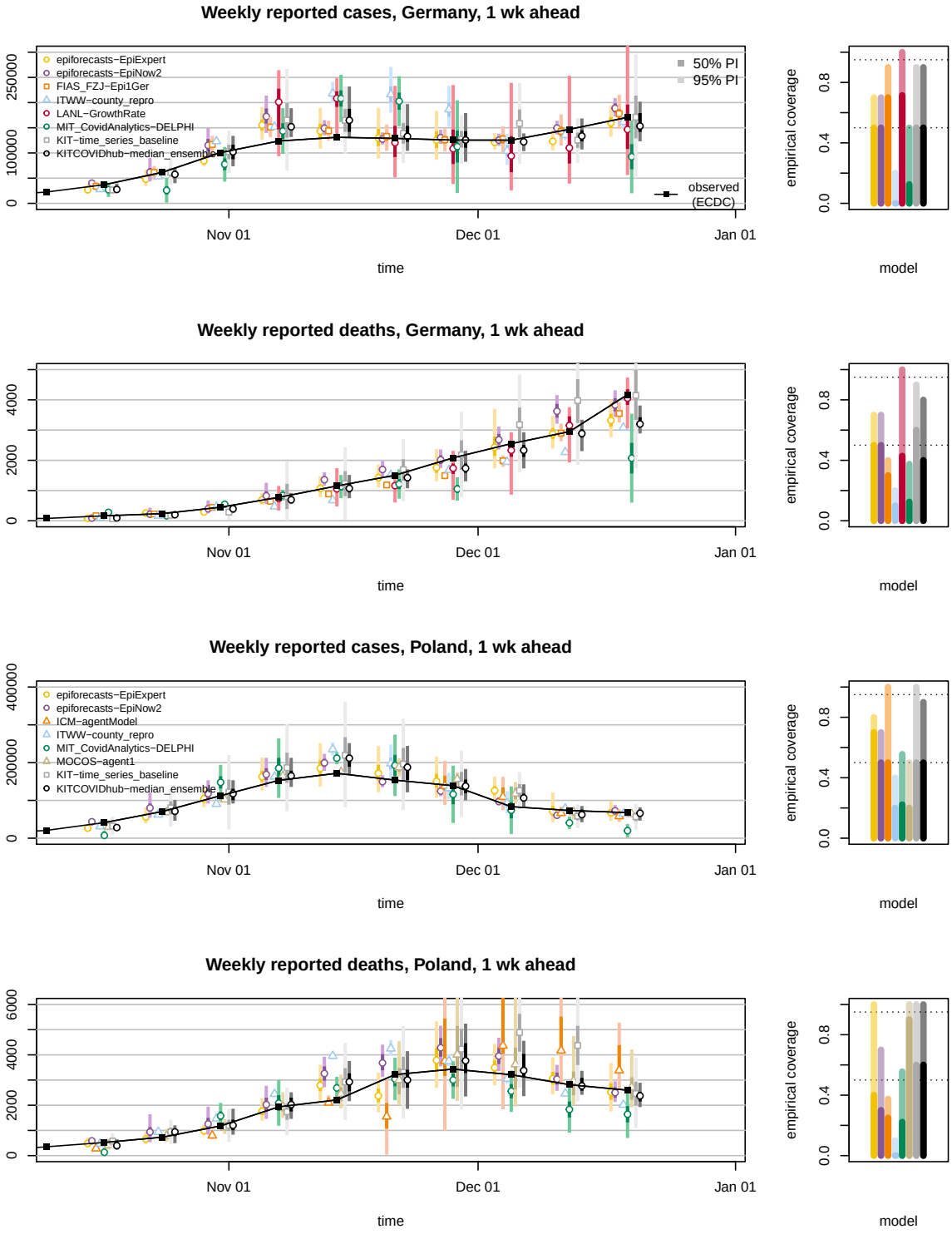

Figure 3: One-week-ahead forecasts of incident cases and deaths in Germany and Poland (left column). Displayed are predictive medians, 50% and 95% prediction intervals. Coverage plots (right column) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

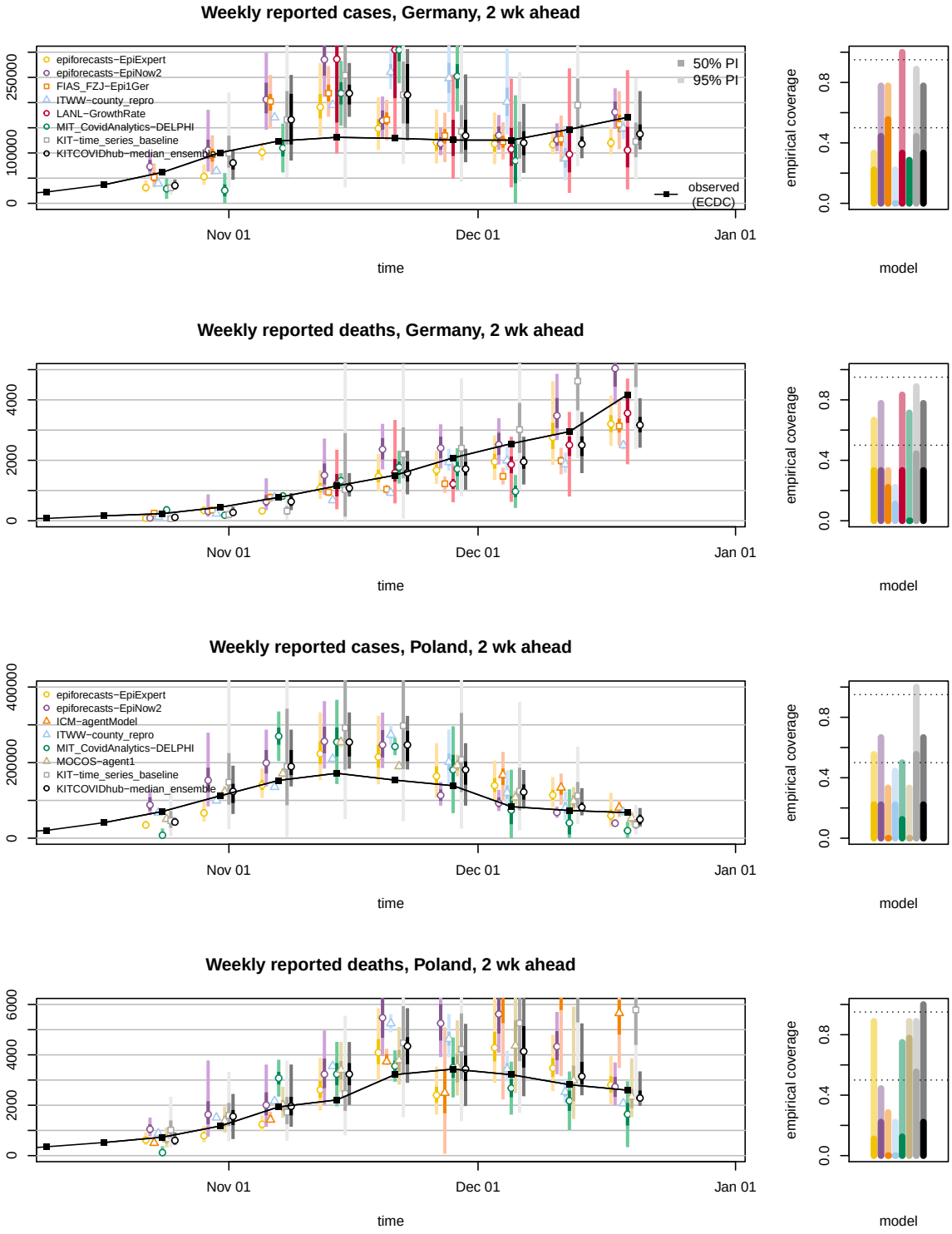

Figure 4: Two-week-ahead forecasts of incident cases and deaths in Germany and Poland (left column). Displayed are predictive medians, 50% and 95% prediction intervals. Coverage plots (right column) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

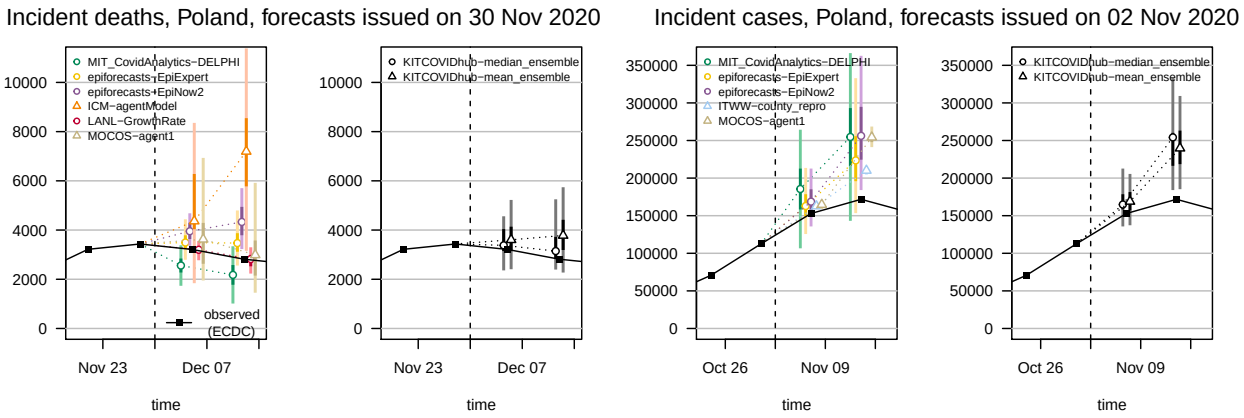

Figure 5: Examples of median and mean ensembles: One- and two-week-ahead forecasts of incident deaths in Poland issued on 30 November (first and second panel), and of incident cases in Poland issued on 2 November 2020 (third and fourth panel). The first and third panels show the member forecasts, the second and fourth panels the respective ensembles. Both predictive medians and 95% (light) and 50% (dark) prediction intervals are shown. The dashed vertical line indicates the date at which the forecasts were issued.

identical. Both distributions are thus rather oddly shaped and not a very plausible belief about the future, with a quarter of the probability mass concentrated in a very short interval. The mean ensemble, on the other hand, produces a more symmetric and thus more realistic representation of the associated uncertainty.

We now briefly address the inverse-WIS ensemble, which is a pragmatic approach to giving more weight to forecasts with good recent performance. Figure 6 shows the weights of the various member models for incident deaths in Germany and Poland. Note that in the first week, numerous models received the same weight as they were submitted for the first time and their scores for past weeks were all imputed with the same values (the worst scores achieved by any model in the respective week). While there are some models which on average receive larger weights than others, weights change considerably over time. Some models are not included in the ensemble for certain weeks, either because of delayed or missing submissions or due to concerns about their plausibility (Section 3.3). The pronounced changes in weights indicate that relative performance fluctuates over time, making it challenging to improve performance of ensemble forecasts by taking past results into account. A possible reason is that models get updated continuously by their maintainers, including major revisions of methodology. Indeed, the overall results shown in Tables 2 and 3 do not indicate any systematic benefits from inverse-WIS weighting.

4.3 Formal forecast evaluation

Forecasts were evaluated using the mean weighted interval score (WIS), mean absolute error (AE) and interval coverage rates. Tables 2 and 3 provide a detailed overview of results by country, target and forecast horizon. We repeated all evaluations using JHU data as ground truth (shown in the Supplement), and the overall results seem robust to this choice. We also provide the same tables for three- and four-week-ahead forecasts in Supplementary Section F, though in view of the discussion in Section 2.2 their usability is limited.

Figure 7 depicts the mean WIS achieved by the different models on the incidence scale. For models providing only point forecasts, the mean AE is shown, which as mentioned in Section 2.3 can be compared to mean WIS values. For deaths, the ensemble forecasts and several submitted models outperform the baseline up to three or even four weeks ahead. As argued before, deaths are a more strongly lagged indicator, which favours predictability at somewhat longer horizons. Another aspect may be that at least in Germany, death numbers have been following a rather uniform upward trend over the study period, making it relatively easy to beat the baseline model. For cases, which are a more immediate measure, almost none of the compared approaches meaningfully outperformed the naïve baseline beyond a horizon of one or two weeks. Especially in Germany this result is largely due to the pronounced overshoot of forecasts in early November

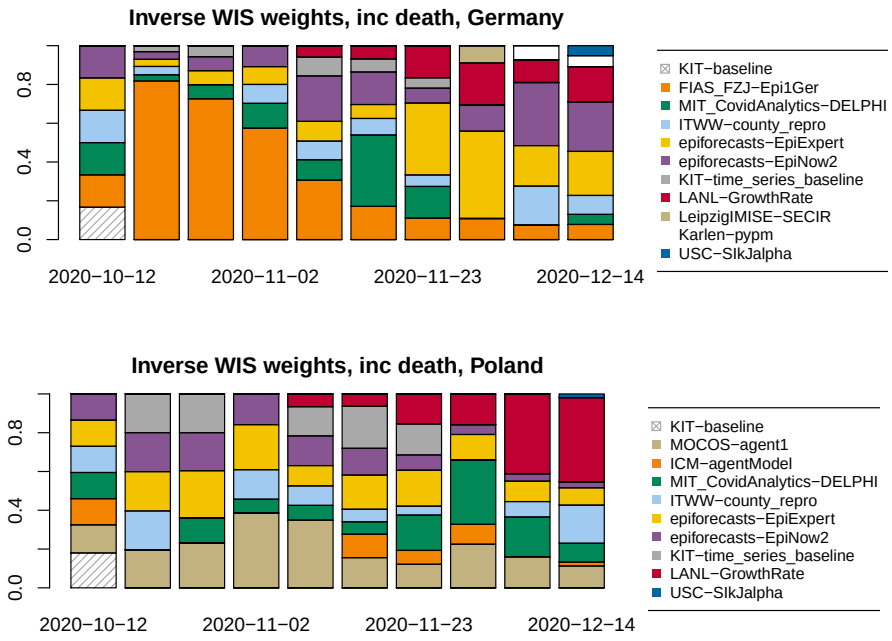

Figure 6: Inverse-WIS weights for forecasts of incident deaths in Germany (top) and Poland (bottom)

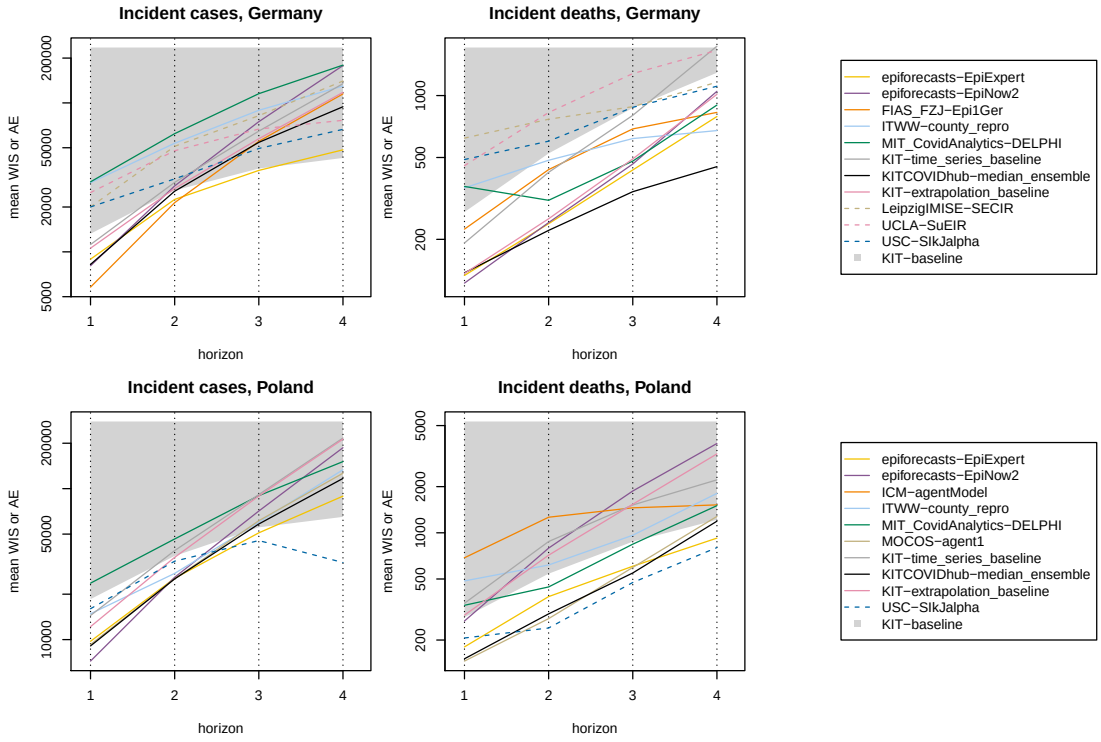

Figure 7: Mean WIS by forecast target and prediction horizon for submitted models and the preregistered median ensemble. For models providing only point forecasts, the mean AE is shown. The lower boundary of the grey area represents the baseline model `KIT-baseline`. Lines crossing the grey area thus indicate that a model fails to outperform the baseline. The numbers underlying this figure can also be found in Tables 2 and 3.

as discussed in Section 4.1. The `KIT-baseline` forecast by definition always predicts a plateau, which is what was observed in Germany for roughly half of the evaluation period. Good performance of the baseline is thus less surprising. Nonetheless, these results underscore that in periods of evolving intervention measures meaningful case forecasts are limited to a rather short time window. In this context we also note that the additional baselines `KIT-extrapolation_baseline` and `KIT-time_series_baseline` do not systematically outperform the naïve baseline and for most targets are neither among the best nor the worst performing approaches.

The median, mean and inverse-WIS ensemble showed overall good, but not outstanding relative performance in terms of mean WIS. Differences between the different ensemble approaches are relatively minor and do not indicate a clear ordering. We re-ran the ensembles retrospectively using all available forecasts, i.e. including those submitted late or excluded due to implausibilities. As can be seen from Supplementary Table 7, this led only to minor changes in performance. Unlike in the US effort (Brooks et al., 2020) the ensemble forecast is not strictly better than the single-model forecasts. Typically, performance is similar to some of the better-performing contributed forecasts, and sometimes the latter have a slight edge (e.g. `FIAS_FZJ-Epi1Ger` for cases in Germany and `MOCOS-agent1` for deaths in Poland). Interestingly, the expert forecast `epiforecasts-EpiExpert` is often among the more successful methods, indicating that an informed human assessment sets a high bar for more formalized model-based approaches. In terms of point forecasts, the extrapolation approach `Geneva-DetGrowth` shows good relative performance, but only covers one-week-ahead forecasts.

The 50% and 95% prediction intervals of most forecasts did not achieve their respective nominal coverage levels (most apparent for cases two weeks ahead). The statistical time series model `KIT-time_series_baseline` features favourably here, though at the expense of wide forecast intervals (Figure 3). While its lack of sharpness leads to mediocre overall performance in terms of the WIS, the model seems to have been a helpful addition to the ensemble by counterbalancing the overconfidence of other models. Indeed, coverage of the 95% intervals of the ensemble is above average, despite not reaching nominal levels.

A last aspect worth mentioning concerns the discrepancies between results for one-week-ahead incident and cumulative quantities. In principle these two should be identical, as forecasts should only be shifted by an additive constant (the last observed cumulative number). This, however, was not the case for all submitted forecasts, and coherence was not enforced by our submission system. For the ensemble forecasts the discrepancies are largely due to the fact that the included models are not always the same.

5 Conclusions

We presented results from a preregistered forecasting project in Germany and Poland, covering 10 weeks during the second wave of the COVID-19 pandemic. We believe that such an effort is helpful to put the outputs from single models in context, and to give a more complete picture of the associated uncertainties. For modelling teams, short-term forecasts can provide a useful feedback loop, via a set of comparable outputs from other models, and regular independent evaluation. A substantial strength of our study is that it took place in the framework of a prespecified evaluation protocol. The criteria for evaluation were communicated in advance, and most considered models covered the entire study period.

Similarly to Funk et al. (2020), we conclude that achieving good predictive accuracy and calibration is challenging in a dynamic epidemic situation. Part of the reason may be that not all models were designed for the sole purpose of short-term forecasting, and could be tailored more specifically to this task. Certain models were originally conceived for what-if projections and retrospective assessments of longer-term dynamics and interventions. This focus on a global fit may limit their flexibility to align closely with the most recent data, making them less successful at short forecast horizons compared to simpler extrapolation approaches. We observed pronounced heterogeneity between the different forecasts, with a general tendency to overconfident forecasting, i.e. too narrow prediction intervals. While over the course of ten weeks, some models showed better average performance (in terms of formal evaluation criteria) than others, relative performance has been fluctuating considerably. Different models may in fact be particularly suitable for different phases of an epidemic (Funk et al., 2020), which is exemplified by the fact that some models were quicker to adjust to slowing growth of cases in Germany. These aspects highlight the importance of considering several independently run models rather than focusing attention on a single one, as is sometimes the case in public

Table 2: Detailed summary of forecast evaluation for Germany (based on ECDC data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 90% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	12,333	8,929	5/10	7/10	30,329	22,497	2/9	3/9	12,334	8,929	5/10	7/10	42,667	31,169	2/9	4/9
epiforecasts-EpiNow2	11,171	8,067	5/10	7/10	37,338	27,712	4/9	7/9	11,171	8,067	5/10	7/10	47,738	34,815	4/9	6/9
FIAS_FZJ-Epi1Ger	7,798	5,796	7/10	9/10	29,190	21,397	5/9	7/9	17,255	11,514	3/10	7/10	38,925	30,312	5/9	6/9
Geneva-DeterministicGrowth	10,963								10,963							
ITWW-county_repro	34,425	29,136	0/10	2/10	64,378	53,604	0/9	2/9	34,077	28,788	0/10	2/10	101,184	84,980	0/9	2/9
LANL-GrowthRate	38,970*	24,029*	5/7	7/7	77,438*	43,758*	2/6	6/6	39,042	27,305	5/10	7/10	116,494	79,819	3/9	5/9
LeipzigIMISE-SECIR	20,019		2/5	3/5	51,115		0/4	1/4	35,901	31,866	1/10	1/10	93,111	84,064	1/9	2/9
MIT_CovidAnalytics-DELPHI	41,313*	29,517*	1/8	4/8	78,872*	62,173*	2/7	2/7								
UCLA-SuEIR	25,012				47,747				25,012				69,800			
USC-SIkJalpha	20,028		1/1	1/1	30,891				21,567		1/1	1/1	49,640			
KIT-baseline	18,475	13,227	5/10	9/10	32,690	25,840	3/9	6/9	18,475	13,227	5/10	9/10	47,472	37,584	3/9	5/9
KIT-extrapolation_baseline	12,016	10,585	7/10	10/10	36,498	26,624	6/9	7/9	12,016	10,585	7/10	10/10	47,145	35,642	7/9	7/9
KIT-time_series_baseline	15,383	11,196	5/10	9/10	44,481	29,286	4/9	8/9	15,383	11,196	5/10	9/10	61,489	40,171	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	14,017	9,552	5/10	9/10	42,063	28,579	2/9	5/9	13,464	9,440	6/10	9/10	52,972	35,935	2/9	8/9
KITCOVIDhub-mean_ensemble	16,649	10,926	4/10	8/10	42,214	27,912	1/9	6/9	15,771	10,844	4/10	8/10	57,125	38,180	1/9	6/9
KITCOVIDhub-median_ensemble	11,534	8,237	5/10	9/10	37,620	25,542	3/9	7/9	12,877	9,394	6/10	7/10	49,438	35,085	2/9	6/9

Germany, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	187	134	5/10	7/10	333	238	3/9	6/9	187	134	5/10	7/10	442	301	3/9	7/9
epiforecasts-EpiNow2	180	123	5/10	7/10	376	241	3/9	7/9	180	123	5/10	7/10	539	345	4/9	7/9
FIAS_FZJ-Epi1Ger	256	224	3/10	4/10	525	437	2/9	3/9	215	176	0/10	4/10	684	569	1/9	4/9
Geneva-DeterministicGrowth	357								357							
Imperial-ensemble2	254	198	5/10	5/10					252	195	5/10	5/10				
ITWW-county_repro	371	356	1/10	2/10	537	485	1/9	2/9	370	355	1/10	2/10	824	760	2/9	2/9
LANL-GrowthRate	195*	131*	3/7	7/7	457*	319*	2/6	5/6	189	136	3/10	7/10	560	446	3/9	5/9
LeipzigIMISE-SECIR	621		0/5	1/5	768		1/4	1/4	1,167	999	0/10	1/10	2,184	1,815	0/9	1/9
MIT_CovidAnalytics-DELPHI	474*	362*	1/8	3/8	403*	310*	0/7	5/7	449*	336*	1/8	3/8	639*	530*	0/7	3/7
UCLA-SuEIR	456				827				456				1,177			
USC-SIkJalpha	489		0/1	0/1	600				500		0/1	0/1	963			
KIT-baseline	479	272	2/10	9/10	835	523	0/9	5/9	479	272	2/10	9/10	1,155	726	0/9	5/9
KIT-extrapolation_baseline	202	137	7/10	9/10	383	252	5/9	8/9	202	137	7/10	9/10	493	337	5/9	8/9
KIT-time_series_baseline	238	192	6/10	9/10	624	424	4/9	8/9	238	192	6/10	9/10	866	589	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	180	117	4/10	9/10	255	152	2/9	8/9	174	111	5/10	9/10	370	230	3/9	8/9
KITCOVIDhub-mean_ensemble	204	141	3/10	9/10	298	179	2/9	8/9	216	149	3/10	9/10	441	269	2/9	8/9
KITCOVIDhub-median_ensemble	200	138	4/10	8/10	334	221	3/9	7/9	202	136	4/10	8/10	440	277	3/9	8/9

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

Table 3: Detailed summary of forecast evaluation for Poland (based on ECDC data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 90% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Poland, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	13,643	9,744	7/10	8/10	37,395	25,497	2/9	5/9	13,620	9,764	7/10	8/10	52,523	34,390	2/9	6/9
epiforecasts-EpiNow2	11,006	7,206	5/10	7/10	38,906	25,875	2/9	6/9	11,028	7,214	5/10	7/10	47,373	31,014	2/9	7/9
Geneva-DeterministicGrowth	7,633								7,656							
ICM-agentModel			2/4	4/4			0/3	1/3			1/4	3/4			0/3	2/3
ITWW-county_repro	18,149	14,831	2/10	4/10	33,298	27,462	2/9	4/9	17,227	13,929	3/10	5/10	50,638	41,512	1/9	4/9
LANL-GrowthRate	15,956*	9,760*	3/7	7/7	49,295*	28,140*	1/6	6/6	15,269	9,559	3/10	10/10	62,801	37,655	2/9	8/9
MIMUW-StochSEIR			3/5	5/5			2/4	2/4			2/5	5/5			1/4	2/4
MIT_CovidAnalytics-DELPHI	32,620*	23,656*	2/9	5/9	60,490*	46,427*	1/8	4/8								
MOCOS-agent1	13,273	9,282	2/10	5/10	31,610	25,243	0/9	3/9	13,273	9,282	2/10	5/10	43,215	32,540	1/9	3/9
USC-SIkJalpha	16,018		0/1	1/1	33,083				19,285		0/1	1/1	50,697			
KIT-baseline	28,164	18,538	5/10	9/10	52,890	35,848	2/9	6/9	28,235	18,574	5/10	9/10	80,765	54,785	2/9	6/9
KIT-extrapolation_baseline	18,311	12,183	6/10	10/10	55,060	35,081	3/9	7/9	18,289	12,178	6/10	10/10	76,607	47,213	3/9	8/9
KIT-time_series_baseline	22,497	14,450	5/10	10/10	60,079	38,901	5/9	9/9	22,475	14,447	5/10	10/10	84,530	53,646	4/9	9/9
KITCOVIDhub-inverse_wis_ensemble	12,768	8,636	4/10	9/10	36,229	25,111	3/9	6/9	11,865	7,905	4/10	9/10	44,477	31,220	2/9	6/9
KITCOVIDhub-mean_ensemble	12,982	8,515	3/10	9/10	36,338	24,129	3/9	6/9	12,051	7,768	5/10	10/10	44,254	30,077	2/9	7/9
KITCOVIDhub-median_ensemble	14,196	9,084	5/10	9/10	39,829	25,254	2/9	6/9	14,033	8,920	5/10	9/10	50,935	31,542	2/9	5/9

Poland, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	285	181	4/10	10/10	605	384	1/9	8/9	285	180	4/10	10/10	874	545	3/9	8/9
epiforecasts-EpiNow2	386	266	3/10	7/10	1,110	795	2/9	4/9	386	266	3/10	7/10	1,528	1,069	2/9	5/9
Geneva-DeterministicGrowth	154								154							
ICM-agentModel	752*	687*	2/8	3/8	1,881*	1,263*	0/7	2/7	1,178*	744*	1/8	4/8	2,955*	1,911*	0/7	3/7
Imperial-ensemble2	397	245	3/10	7/10					369	217	3/10	8/10				
ITWW-county_repro	525	486	0/10	1/10	701	617	0/9	2/9	524	485	0/10	1/10	1,219	1,091	0/9	2/9
LANL-GrowthRate	239*	178*	4/7	7/7	404*	257*	3/6	6/6	216	154	5/10	10/10	637	413	3/9	8/9
MIMUW-StochSEIR			1/5	4/5			1/4	2/4			1/5	4/5			0/4	4/4
MIT_CovidAnalytics-DELPHI	512*	337*	2/9	5/9	663*	444*	1/8	6/8	597*	424*	2/9	6/9	1,075*	795*	1/8	3/8
MOCOS-agent1	194	146	9/10	10/10	420	276	7/9	8/9	194	146	9/10	10/10	556	386	7/9	9/9
USC-SIkJalpha	206		0/1	1/1	240				256		0/1	0/1	242			
KIT-baseline	437	281	5/10	10/10	834	541	2/9	7/9	437	281	5/10	10/10	1,245	812	2/9	6/9
KIT-extrapolation_baseline	408	291	6/10	8/10	996	715	5/9	7/9	408	291	6/10	8/10	1,423	1,007	4/9	7/9
KIT-time_series_baseline	546	347	6/10	10/10	1,371	877	5/9	8/9	546	347	6/10	10/10	1,921	1,242	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	220	156	6/10	10/10	488	320	4/9	8/9	242	166	7/10	10/10	702	460	4/9	9/9
KITCOVIDhub-mean_ensemble	252	166	7/10	9/10	585	372	4/9	8/9	265	175	7/10	9/10	815	534	4/9	9/9
KITCOVIDhub-median_ensemble	215	151	6/10	10/10	471	296	2/9	9/9	231	163	6/10	10/10	707	469	4/9	9/9

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

discussions. Here, collaborative forecasting projects can provide valuable insights. Overall, ensemble methods showed good, but not outstanding relative performance, notably with clearly above-average coverage rates. Its improved reliability is a key strength of the ensemble approach, and we expect that the continuing refinement of member models will further strengthen the robustness of the ensemble. An important question is whether ensemble forecasts could be improved by sensible weighting of members or post-processing steps. Given the limited amount of available forecast history and rapid changes in the epidemic situation, this is a challenging encounter, and indeed we did not find benefits in the inverse-WIS approach.

An obvious extension to both assess forecasts in more detail and make them more relevant to decision makers is to issue them at a finer geographical resolution. During the evaluation period covered in this work, only three of the contributed forecast models (`ITWW-county_repro` and `USC-SIkJalpha`, `LeipzigIMISE-SECIR` for the state of Saxony) also provided forecasts at the regional level (German states, Polish voivodeships). Extending this to a larger number of models is one of the main priorities for the further course of the German and Polish Forecast Hub project.

In its present form, the project covers only forecasts of confirmed cases and deaths. These commonly addressed forecasting targets were already covered by a critical mass of teams when the project was started. Given limited available time resources of teams, a choice was made to focus efforts on this narrow set of targets. An extension to other quantities such as hospitalizations or ICU/ventilation need, which have important public health implications, was considered, but in view of emerging parallel efforts and open questions on data availability not prioritized.

The German and Polish Forecast Hub will continue to compile short-term forecasts and process them into forecast ensembles. With vaccine rollout likely to start in early 2021, models will face a new layer of complexity. We aim to provide further systematic evaluations for these future phases, contributing to a growing body of evidence on the potential and limits of pandemic short-term forecasting.

Reproducibility / data availability

All data used in this article are publicly available at https://github.com/KITmetricslab/covid19-forecast-hub-de. Forecasts can be visualized interactively at https://github.com/KITmetricslab/covid19-forecast-hub-de. Codes to reproduce figures and tables are available at https://github.com/KITmetricslab/analyses_de_pl.

References

Abbott, S., Hellewell, J., Sherratt, K., Gostic, K., Hickson, J., Badr, H. S., DeWitt, M., Thompson, R., and Funk, S. (2020). epiforecasts/EpiNow2: Prerelease. Available at https://zenodo.org/record/4343617 (last accessed 22 December 2020).

Adamik, B., Bawiec, M., Bezborodov, V., Bock, W., Bodych, M., Burgard, J. P., Götz, T., Krueger, T., Migalska, A., Pabjan, B., Ożański, T., Rafajłowicz, E., Rafajłowicz, W., Skubalska-Rafajłowicz, E., Ryfczyńska, S., Szczurek, E., and Szymański, P. (2020). Mitigation and herd immunity strategy for covid-19 is likely to fail. *medRxiv*, https://www.medrxiv.org/content/10.1101/2020.03.25.20043109v1.

Afelt, A., Bartczuk, R., Biecek, P., Bodych, M., Gambin, A., Gogolewski, K., Kaczorek, A., Kisielewski, J., Krüger, T., Migalska, Mikołajczyk, R., Moszyński, A., Niedzielewski, K., Nowosielski, A., Pabjan, B., Radwan, M., Rakowski, F., Rosińska, M., Semeniuk, M., and Zieliński, J. (2020). Quo vadis coronavirus? Rekomendacje zespołów epidemiologii obliczeniowej na rok 2021. Available online at https://quovadis.crs19.pl/ (last accessed 22 December 2020).

Barbarossa, M. V., Fuhrmann, J., Meinke, J. H., Krieg, S., Varma, H. V., Castelletti, N., and Lippert, T. (2020). Modeling the spread of COVID-19 in Germany: Early assessment and possible scenarios. *PLOS ONE*, 15(9):e0238559.

Bicher, M., Zuba, M., Rainer, L., Bachner, F., Rippinger, C., Ostermann, H., Popper, N., Thurner, S., and Klimek, P. (2020). Supporting Austria through the COVID-19 epidemics with a forecast-based early warning system. *medRxiv*, https://www.medrxiv.org/content/10.1101/2020.10.18.20214767v2.

Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2020a). Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, in press.

Bracher, J., the German and Polish COVID-19 Forecast Hub Team, and Participants (2020b). Study protocol: Comparison and combination of real-time COVID19 forecasts in Germany and Poland. Deposited 8 October 2020, Registry of the Open Science Foundation, https://osf.io/k8d39.

Brooks, L. C., Ray, E. L., Bien, J., Bracher, J., Rumack, A., Tibshirani, R. J., and Reich, N. G. (2020). Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the U.S. Blog entry, International Institute of Forecasters, https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/, full paper to follow.

COVID-19 Forecast Hub Team (2020). COVID-19 Forecast Hub – Projections of COVID-19, in standardized format. Available online at https://github.com/reichlab/covid19-forecast-hub and https://covid19forecasthub.org/.

Dean, N. E., Pastore y Piontti, A., Madewell, Z. J., Cummings, D. A., Hitchings, M. D., Joshi, K., Kahn, R., Vespignani, A., Halloran, M. E., and Longini, I. M. (2020). Ensemble forecast modeling for the design of COVID-19 vaccine efficacy trials. *Vaccine*, 38(46):7213–7216.

Del Valle, S. Y., McMahon, B. H., Asher, J., Hatchett, R., Lega, J. C., Brown, H. E., Leany, M. E., Pantazis, Y., Roberts, D. J., Moore, S., Peterson, A. T., Escobar, L. E., Qiao, H., Hengartner, N. W., and Mukundan, H. (2018). Summary results of the 2014–2015 DARPA Chikungunya Challenge. *BMC Infectious Diseases*, 18(1):245.

Desai, A. N., Kraemer, M. U. G., Bhatia, S., Cori, A., Nouvellet, P., Herringer, M., Cohn, E. L., Carrion, M., Brownstein, J. S., Madoff, L. C., and Lassmann, B. (2019). Real-time epidemic forecasting: Challenges and opportunities. *Health Security*, 17(4):268–275. PMID: 31433279.

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

European Centre for Disease Prevention and Control (2020a). Baseline projections of COVID-19 in the EU/EEA and the UK: Update. Published 17 September 2020, https://www.ecdc.europa.eu/sites/default/files/documents/ECDC-30-day-projections-Sept-2020.pdf.

European Centre for Disease Prevention and Control (2020b). Download historical data (to 14 December 2020) on the daily number of new reported COVID-19 cases and deaths worldwide. Available online at https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

European Centre for Disease Prevention and Control (2020c). Projected baselines of COVID-19 in the EU/EEA and the UK for assessing the impact of de-escalation of measures. Published 26 May 2020, https://www.ecdc.europa.eu/sites/default/files/documents/Projected-baselines-COVID-19-for-assessing-impact-measures.pdf.

Funk, S., Abbott, S., Atkins, B., Baguelin, M., Baillie, J., Birrell, P., Blake, J., Bosse, N., Burton, J., Carruthers, J., Davies, N., De Angelis, D., Dyson, L., Edmunds, W., Eggo, R., Ferguson, N., Gaythorpe, K., Gorsich, E., Guyver-Fletcher, G., Hellewell, J., Hill, E., Holmes, A., House, T., Jewell, C., Jit, M., Jombart, T., Joshi, I., Keeling, M., Kendall, E., Knock, E., Kucharski, A., Lythgoe, K., Meakin, S., Munday, J., Openshaw, P., Overton, C., Pagani, F., Pearson, J., Perez-Guzman, P., Pellis, L., Scarabel, F., Semple, M., Sherratt, K., Tang, M., Tildesley, M., Van Leeuwen, E., Whittles, L., CMMID COVID-19 Working Group, Imperial College COVID-19 Response Team, and ISARIC4C Investigators (2020). Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. *medRxiv*, https://www.medrxiv.org/content/10.1101/2020.11.11.20220962v2.

Funk, S., Camacho, A., Kucharski, A. J., Lowe, R., Eggo, R. M., and Edmunds, W. J. (2019). 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):1–17.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, 310(5746):248–249.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Günther, F., Bender, A., Katz, K., Küchenhoff, H., and Höhle, M. (2020). Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal*, https://doi.org/10.1002/bimj.202000112.

Held, L., Meyer, S., and Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: The 13th Armitage lecture. *Statistics in Medicine*, 36(22):3443–3460.

Hyndman, R. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22.

Johansson, M. A., Apfeldorf, K. M., Dobson, S., Devita, J., Buczak, A. L., Baugher, B., Moniz, L. J., Bagley, T., Babin, S. M., Guven, E., Yamana, T. K., Shaman, J., Moschou, T., Lothian, N., Lane, A., Osborne, G., Jiang, G., Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., Rosenfeld, R., Lessler, J., Reich, N. G., Cummings, D. A. T., Lauer, S. A., Moore, S. M., Clapham, H. E., Lowe, R., Bailey, T. C., García-Díez, M., Carvalho, M. S., Rodó, X., Sardar, T., Paul, R., Ray, E. L., Sakrejda, K., Brown, A. C., Meng, X., Osoba, O., Vardavas, R., Manheim, D., Moore, M., Rao, D. M., Porco, T. C., Ackley, S., Liu, F., Worden, L., Convertino, M., Liu, Y., Reddy, A., Ortiz, E., Rivero, J., Brito, H., Juarrero, A., Johnson, L. R., Gramacy, R. B., Cohen, J. M., Mordecai, E. A., Murdock, C. C., Rohr, J. R., Ryan, S. J., Stewart-Ibarra, A. M., Weikel, D. P., Jutla, A., Khan, R., Poultney, M., Colwell, R. R., Rivera-García, B., Barker, C. M., Bell, J. E., Biggerstaff, M., Swerdlow, D., Mier-y Teran-Romero, L., Forshey, B. M., Trtanj, J., Asher, J., Clay, M., Margolis, H. S., Hebbeler, A. M., George, D., and Chretien, J.-P. (2019). An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274.

Keeling, M. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton, NJ.

Li, M. L., Tazi Bouardi, H., Skali Lami, O., Trikalinos, T. A., Trichakis, N. K., and Bertsimas, D. (2020). Forecasting COVID-19 and analyzing the effect of government interventions. *medRxiv*, https://www.medrxiv.org/content/10.1101/2020.06.23.20138693v1.

Lichtendahl, K. C., Grushka-Cockayne, Y., and Winkler, R. L. (2013). Is it better to average probabilities or quantiles? *Management Science*, 59(7):1594–1611.

Nature Publishing Group (2020). Editorial: Developing infectious disease surveillance systems. *Nature Communications*, 11:4962.

Petropoulos, F. and Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. *PLOS ONE*, 15(3):e0231236.

Rakowski, F., Gruziel, M., Bieniasz-Krzywiec, L., and Radomski, J. P. (2010). Influenza epidemic spread simulation for Poland – a large scale, individual based model study. *Physica A: Statistical Mechanics and its Applications*, 389(16):3149–3165.

Ray, E. L., Wattanachit, N., Niemi, J., Kanji, A. H., House, K., Cramer, E. Y., Bracher, J., Zheng, A., Yamana, T. K., Xiong, X., Woody, S., Wang, Y., Wang, L., Walraven, R. L., Tomar, V., Sherratt, K., Sheldon, D., Reiner, R. C., Prakash, B. A., Osthus, D., Li, M. L., Lee, E. C., Koyluoglu, U., Keskinocak, P., Gu, Y., Gu, Q., George, G. E., España, G., Corsetti, S., Chhatwal, J., Cavany, S., Biegel, H., Ben-Nun, M., Walker, J., Slayton, R., Lopez, V., Biggerstaff, M., Johansson, M. A., Reich, N. G., and COVID-19 Forecast Hub Consortium (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. *medRxiv*, https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

Reich, N. and Rivers, C. (2020). Scientists want to predict COVID-19's long-term trajectory. Here's why they can't. *Washington Post*. Published 15 September 2020, https://www.washingtonpost.com/outlook/2020/09/15/scientists-want-predict-covid-19s-long-term-trajectory-heres-why-they-cant/.

Reich, N. G., Brooks, L. C., Fox, S. J., Kandula, S., McGowan, C. J., Moore, E., Osthus, D., Ray, E. L., Tushar, A., Yamana, T. K., Biggerstaff, M., Johansson, M. A., Rosenfeld, R., and Shaman, J. (2019a). A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154.

Reich, N. G., McGowan, C. J., Yamana, T. K., Tushar, A., Ray, E. L., Osthus, D., Kandula, S., Brooks, L. C., Crawford-Crudell, W., Gibson, G. C., Moore, E., Silva, R., Biggerstaff, M., Johansson, M. A., Rosenfeld, R., and Shaman, J. (2019b). Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLOS Computational Biology*, 15(11):e1007486.

Robert Koch Institute (2020). Coronavirus disease 2019 – daily situation report of the Robert Koch Institute, 18 November 2020. Published at https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Nov_2020/2020-11-18-en.pdf.

Srivastava, A., Xu, T., and Prasanna, V. K. (2020). Fast and accurate forecasting of COVID-19 deaths using the SIkJα model. *arXiv*, https://arxiv.org/abs/2007.05180.

Taylor, K. S. and Taylor, J. W. (2020). A comparison of aggregation methods for probabilistic forecasts of COVID-19 mortality in the United States. *arXiv*, https://arxiv.org/abs/2007.11103.

UK Scientific Pandemic Influenza Group on Modelling (2020). Medium-term projections and model descriptions. Consensus statement, considered at UK SAGE 66 on 5 November 2020, https://www.gov.uk/government/publications/spi-m-o-covid-19-medium-term-projections-explainer-31-october-2020.

Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., Zhang, Q., Chowell, G., Simonsen, L., Vespignani, A., and the RAPIDD Ebola Forecasting Challenge group (2018). The RAPIDD Ebola Forecasting Challenge: Synthesis and lessons learnt. *Epidemics*, 22:13–21.

Viboud, C. and Vespignani, A. (2019). The future of influenza forecasts. *Proceedings of the National Academy of Sciences*, 116(8):2802–2804.

Yamana, T. K., Kandula, S., and Shaman, J. (2016). Superensemble forecasts of dengue outbreaks. *Journal of The Royal Society Interface*, 13(123):20160410.

Zamo, M., Bel, L., and Mestre, O. (2020). Sequential aggregation of probabilistic forecasts—application to wind speed ensemble forecasts. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, in press.

Zou, D., Wang, L., Xu, P., Chen, J., Zhang, W., and Gu, Q. (2020). Epidemic model guided machine learning for COVID-19 forecasts in the United States. *medRxiv*, https://www.medrxiv.org/content/10.1101/2020.05.24.20111989v1.

Acknowledgements

We are grateful to the team of the US COVID-19 Forecast Hub, in particular Evan L. Ray and Nicholas G. Reich, for fruitful exchange and their support. We would like to thank Dean Karlen for contributions to the Forecast Hub from December 2020 onwards and Berit Lange for helpful comments. We moreover want to thank Fabian Eckelmann and Knut Persecke who implemented the interactive visualization tool.

The work of Johannes Bracher was supported by the Helmholtz Foundation via the SIMCARD Information and Data Science Pilot Project. Sangeeta Bhatia acknowledges support from the Wellcome Trust (219415). Nikos I. Bosse acknowledges funding by the Health Protection Research Unit (grant code NIHR200908). Sebastian Funk and Sam Abbott acknowledge support from the Wellcome Trust (grant no.

210758/Z/18/Z). The work of Ajitesh Srivastava was supported by National Science Foundation Award No. 2027007 (RAPID). Tilmann Gneiting and Daniel Wolffram are grateful for support by the Klaus Tschira Foundation.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the institutions they are affiliated with.

Author contributions

JB, TG and MS conceived the study with advice from AU. JB, DW, JD, KG and JK put in place and maintained the forecast submission and processing system. AU coordinated the creation of an interactive visualization tool. JB performed the evaluation analyses with inputs from DW, TG, MS and members of various teams. SA, MVB, DB, SB, MB, NIB, JPB, LC, GF, JF, SF, KG, QG, SH, TH, YK, HK, TK, EK, MLL, JHM, IJM, KN, TO, FR, MS, SS, AS, JZ and DZ contributed forecasts (see list of contributors by team). JB, TG and MS wrote the manuscript, with TK, MB and KG contributing to the description of intervention measures in Poland. All teams and members of the coordinating team provided feedback on the manuscript and descriptions of the respective models.

Competing interests

The authors declare no competing interests.

List of contributors by team

`CovidAnalytics-DELPHI`: Michael Lingzhi Li (Operations Research Center, Massachusetts Institute of Technology, Cambridr, MA, USA), Dimitris Bertsimas, Hamza Tazi Bouardi, Omar Skali Lami, Saksham Soni (all Sloan School of Management, Massachusetts Institute of Technology, USA)

`epiforecasts-EpiExpert` and `epiforecasts-EpiNow2`: Sam Abbott, Nikos I. Bosse, Sebastian Funk (all London School of Hygiene and Tropical Medicine, London, UK)

`FIAS_FZJ-Epi1Ger`: Maria Vittoria Barbarossa (Frankfurt Institute for Advanced Studies, Frankfurt, Germany), Jan Fuhrmann (Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany and Frankfurt Institute for Advanced Studies, Frankfurt, Germany), Jan H. Meinke (Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany)

`Geneva-DeterministicGrowth`: Antoine Flahault, Elisa Manetti (both Institute of Global Health, Faculty of Medicine, University of Geneva, Geneva, Switzerland), Christine Choirat, Benjamin Bejar Haro, Ekaterina Krymova, Gavin Lee, Guillaume Obozinski, Tao Sun (all Swiss Data Science Center, ETH Zurich and EPFL Lausanne, Switzerland), Dorina Thanou (Center for Intelligent Systems, EPFL, Lausanne Switzerland)

`ICM-agentModel`: Łukasz Górski, Magdalena Gruziel-Słomka, Artur Kaczorek, Antoni Moszyński, Karol Niedzielewski, Jedrzej Nowosielski, Maciej Radwan, Franciszek Rakowski, Marcin Semeniuk, Jakub Zieliński (all Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland), Rafał Bartczuk (Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw and Institute of Psychology, John Paul II Catholic University of Lublin, Lublin, Poland), Jan Kisielewski (Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw and Faculty of Physics, University of Białystok, Ciołkowskiego 1L, 15-245 Białystok)

`Imperial-ensemble2`: Sangeeta Bhatia (MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), Imperial College, London, UK)

`ITWW-county_repro`: Przemyslaw Biecek (Warsaw University of Technology, Warsaw, Poland), Viktor Bezborodov, Marcin Bodych, Tyll Krueger (all Wroclaw University of Science and Technology, Poland), Jan Pablo Burgard (Economic and Social Statistics Department, University of Trier, Germany), Stefan Heyder, Thomas Hotz (both Institute of Mathematics, Technische Universität Ilmenau, Germany)

`LeipzigIMISE-SECIR`: Yuri Kheifetz, Holger Kirsten, Markus Scholz (all Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany)

`MIMUW-StochSEIR`: Anna Gambin, Krzysztof Gogolewski, Błażej Miasojedow, Ewa Szczurek (all Institute of Informatics, University of Warsaw, Warsaw, Poland), Daniel Rabczenko, Magdalena Rosińska (Polish National Institute of Public Health – National Institute of Hygiene)

`MOCOS-agent1`: Marek Bawiec, Viktor Bezborodov, Marcin Bodych, Tyll Krueger, Tomasz Ożański, Barbara Pabjan, Ewaryst Rafajłowicz, Ewa Skubalska-Rafajłowicz, Wojciech Rafajłowicz (all Wroclaw University of Science and Technology, Poland), Przemyslaw Biecek (Warsaw University of Technology), Agata Migalska (Wroclaw University of Science and Technology, Poland and Nokia Solutions and Networks, Wroclaw, Poland), Ewa Szczurek (University of Warsaw)

`UCLA-SuEIR`: Quanquan Gu, Pan Xu, Jinghui Chen, Lingxiao Wang, Difan Zou, Weitong Zhang (all Department of Computer Science, University of California, Los Angeles, USA)

`USC-SIkJalpha` Ajitesh Srivastava, Viktor K. Prasanna, Frost Tianjian Xu (all University of Southern California, Los Angeles, USA)

Supplementary Materials for Bracher et al (2020): Short-term forecasting of COVID-19 in Germany and Polandduring the second wave – a preregistered study

A Stratified visualizations of case and death counts

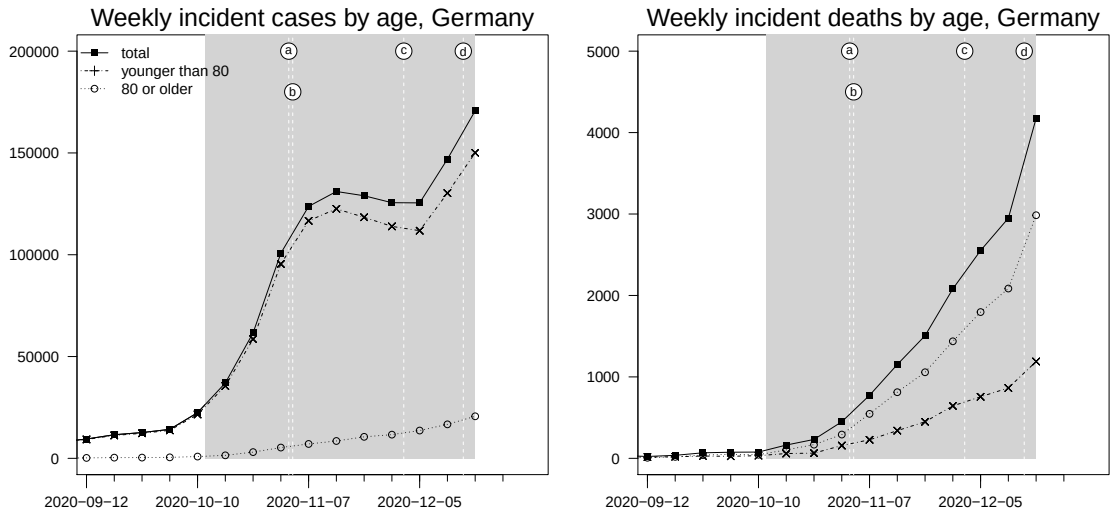

Figure 8: Weekly incident COVID-19 cases and deaths in Germany, pooled and stratified by age below and above 80 years. Events marked by letters a – d are explained in Figure 1.

B Detailed description of baseline forecasts

We here describe the three baseline forecasts from Section 3.1 in more detail.

B.1 `KIT-baseline`

Denote the quantity of interest on the incidence scale by X_t. The corresponding quantity on the cumulative scale is denoted by $Y_t = \sum_{s \leq t} X_t$. The one-week-ahead forecast for X_{t+1} is given by a negative binomial distribution with mean X_t and overdispersion parameter ψ. Due to the skewness of the negative binomial distribution this implies that the predictive median is slightly smaller than X_t. The overdispersion parameter is estimated from the last five available observations using a maximum likelihood approach, i.e. by maximizing

$$\sum_{i=0}^{4} \log \pi(X_{t-i} \mid X_{t-i-1}, \psi),$$

with respect to ψ, where $\pi(\cdot \mid X_{t-i-1}, \psi)$ is the probability mass function of a negative binomial distribution with mean X_{t-i-1} and overdispersion parameter ψ. For technical reasons we replace any mean of a negative binomial distribution which would equal zero by 0.2. The two- to four-week-ahead forecasts are simply set to the same distribution as the one-week-ahead forecast.

To obtain forecasts on the cumulative scale we assume independence between $X_{t+1}, X_{t+2}, X_{t+3}$ and X_{t+4}. As the sum of independent random variables following negative binomial distributions with the same overdispersion parameter follows again a negative binomial distribution, $Y_{t+1}, Y_{t+2}, Y_{t+3}$ and Y_{t+4} follow shifted negative binomial distributions with overdispersion parameter $\psi, 2\psi, 3\psi$ and 4ψ, respectively.

B.2 `KIT-extrapolation_baseline`

We assume again a (conditional) negative binomial distribution, but with mean $\lambda_{t+1} = \alpha X_t$ rather than just X_t. The parameter α is estimated from the last three observed values in the following way:

- If the last three observations are ordered, i.e. $X_{t-2} < X_{t-1} < X_t$ or $X_{t-2} > X_{t-1} > X_t$ we let

  $$\alpha = \frac{X_t}{X_{t-1}},$$

  which corresponds to simple multiplicative extrapolation.
- Otherwise we let $\alpha = 1$, so that the predictive mean λ_{t+1} equals the last observation X_t.

The idea behind this distinction is that the model should only use trends if they have manifested for at least two weeks. The overdispersion parameter is estimated by maximizing

$$\sum_{i=1}^{5} \log \pi(X_{t-i} \mid \lambda_{t-i}, \psi),$$

with respect to ψ (keeping the value α entering into $\lambda_{t-i} = \alpha X_{t-i-1}$ constant at the value chosen as described above). Note that we do not use the last observation X_t here as by construction (if the last three observations are ordered) $X_t = \lambda_t$.

We then sample 100,000 paths $(X_{t+1}, X_{t+2}, X_{t+3}, X_{t+4})$ from this model and obtain forecast quantiles for both incident and cumulative quantities from these samples.

B.3 `KIT-time_series_baseline`

We fit an exponential smoothing model with multiplicative errors and without seasonality to the last 12 observations on the incidence scale. The `R` (R Core Team, 2020) command is

```
forecast::ets(ts, model="MMN")
```

using the `forecast` package (Hyndman and Khandakar, 2008). As noted in the main text, this specification is taken from Petropoulos and Makridakis (2020). As in the previous section we proceed by sampling paths from this model and computing predictive quantiles from them.

C Sources on changes in non-pharmaceutical interventions and testing regimes

We here provide sources for the dates of interventions shown in Figure 1.

Poland: Government interventions are largely documented on the respective governmental web site and the Twitter channel of the Polish Ministry of Health (in Polish):

- https://www.gov.pl/web/koronawirus/100-dni-solidarnosci-w-walce-z-covid-19
- https://twitter.com/MZ_GOV_PL.

Specific news items on mentioned interventions/events:

- Four symptoms required for test: Ministerstwo Zdrowia przekazało zasady zlecania testów na koronawirusa, Wprost, 23 September 2020, https://www.wprost.pl/koronawirus-w-polsce/10368723/ministerstwo-zdrowia-przekazalo-zasady-zlecania-testow-na-koronawirusa.html (last accessed 22 December 2020).
- Only one out of four symptoms required for test: Dlaczego lekarz odmawia skierowania na test na COVID-19? Medonet, 5 Nov 2020, https://www.medonet.pl/koronawirus/koronawirus-w-polsce,kiedy-lekarz-moze-odmowic-skierowania-na-test-na-koronawirusa,artykul,26303647.html (last accessed 22 December 2020)
- Bulk reporting of 22,000 cases on 25 November: Rozbieżności w statystykach koronawirusa. 22 tys. przypadków będą doliczone do ogólnej liczby wyników, Forsal, 23 November, https://forsal.pl/lifestyle/zdrowie/artykuly/8017628,rozbieznosci-w-statystykach-koronawirusa-22-tys-przypadkow-beda-doliczone-do-ogolnej-liczby-wynikow.html (last accessed 22 December 2020)
- High test positivity and suspected under-ascertainment: Polish doctors fear high rate of positive COVID tests show pandemic worse than it appears, J. Plucinska, Reuters, 1 December 2020, https://www.reuters.com/article/us-health-coronavirus-poland-cases/polish-doctors-fear-high-rate-of-positive-covid-tests-show-pandemic-worse-than-it-appears-idUSKBN28B54Q (last accessed 22 December 2020)

Germany: A chronicle of the most important events (in German) can be found on the web site of the Germany Ministry of Health:

- https://www.bundesgesundheitsministerium.de/coronavirus/chronik-coronavirus.html

Specific news items on mentioned interventions/events:

- New testing strategy announced: SARS-CoV-2-Diagnostik: RKI passt Testempfehlungen an, Ärzteblatt, 3 November 2020, https://www.aerzteblatt.de/nachrichten/118001/SARS-CoV-2-Diagnostik-RKI-passt-Testempfehlungen-an (last accessed 22 December 2020)
- Semi-lockdown from 2 November onwards: Coronavirus: Germany to impose one-month partial lockdown, Deutsche Welle, 28 October 2020, https://www.dw.com/en/coronavirus-germany-to-impose-one-month-partial-lockdown/a-55421241 (last accessed 22 December 2020)
- Reinforced rules from 1 December onwards: Was gilt wo im Corona-Dezember? Tagesschau, 1 December 2020, https://www.tagesschau.de/inland/corona-plan-bundeslaender-beschluss-103.html (last accessed 22 December 2020)
- Full lockdown starting on 16 December: Lockdown in Deutschland – Das sind die Corona-Regeln. Tagesschau, 13 December 2020, https://www.tagesschau.de/inland/corona-regeln-lockdown-101.html (last accessed 22 December 2020)

D Availability and delays of forecasts

Table 4: Availability of forecasts by model, target and forecast horizon.

Each entry describes up to which forecast horizon (in weeks) forecasts for incident cases, cumulative cases, incident deaths and cumulative deaths were made available (numbers in this order and separated by semicolons). Asterisks indicate that forecasts were only available on Wednesday or later rather than before Tuesday 3pm.

Germany

	2020-10-12	2020-10-19	2020-10-26	2020-11-02	2020-11-09	2020-11-16	2020-11-23	2020-11-30	2020-12-07	2020-12-14
epiforecasts-EpiExpert	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
epiforecasts-EpiNow2	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
FIAS_FZJ-Epi1Ger	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
Geneva-DeterministicGrowth	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1*
Imperial-ensemble2	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1
ITWW-county_repro	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4
KIT-baseline	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
KIT-extrapolation_baseline	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4
KIT-time_series_baseline	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*
inverse_wis_ensemble	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*
mean_ensemble	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
median_ensemble	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
LANL-GrowthRate	-; 4; -; 4	-; 4; -; 4	-; 4; -; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
LeipzigIMISE-SECIR	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*
MIT_CovidAnalytics-DELPHI	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	-; -; -; -	-; -; -; -	4; -; 4; 4
UCLA-SuEIR	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*
USC-SIkJalpha	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4

Poland

	2020-10-12	2020-10-19	2020-10-26	2020-11-02	2020-11-09	2020-11-16	2020-11-23	2020-11-30	2020-12-07	2020-12-14
epiforecasts-EpiExpert	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
epiforecasts-EpiNow2	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
Geneva-DeterministicGrowth	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1	1; 1; 1; 1*
ICM-agentModel	-; -; 4; 4	-; -; -; -	-; -; 4; 4*	-; -; -; -	-; -; 4; 4	-; -; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4
Imperial-ensemble2	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1*	-; -; 1; 1
ITWW-county_repro	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4
KIT-baseline	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
KIT-extrapolation_baseline	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4
KIT-time_series_baseline	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
inverse_wis_ensemble	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
mean_ensemble	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
median_ensemble	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
LANL-GrowthRate	-; 4; -; 4	-; 4; -; 4	-; 4; -; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
MIMUW-StochSEIR	-; -; -; -	-; -; -; -	-; -; -; -	-; -; -; -	-; -; -; -	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4*	4; 4; 4; 4
MIT_CovidAnalytics-DELPHI	4; -; 4; 4	-; -; -; -	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4	4; -; 4; 4
MOCOS-agent1	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4*	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4
USC-SIkJalpha	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4	4; 4; 4; 4

Most forecasts from Imperial-ensemble2 were only made available retrospectively to the Forecast Hub, but had previously been shown in real time on the web dashboard of the Imperial team.

E Additional results for one- and two-week-ahead forecasts

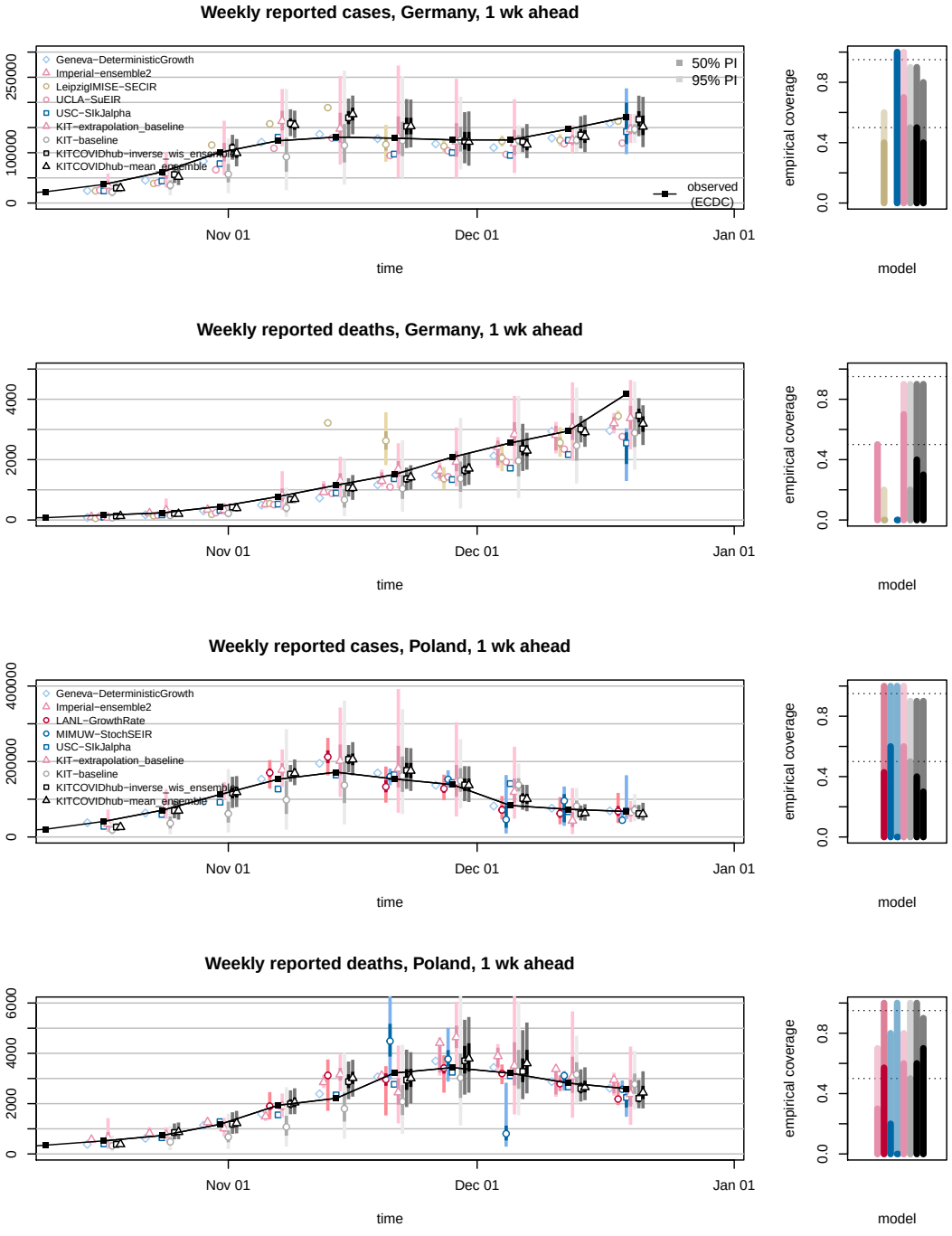

Figure 9: One-week-ahead forecasts of incident cases and deaths in Germany and Poland (left). Displayed are predictive medians, 50% and 95% prediction intervals for models not shown in Figure 3. Coverage plots (right) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

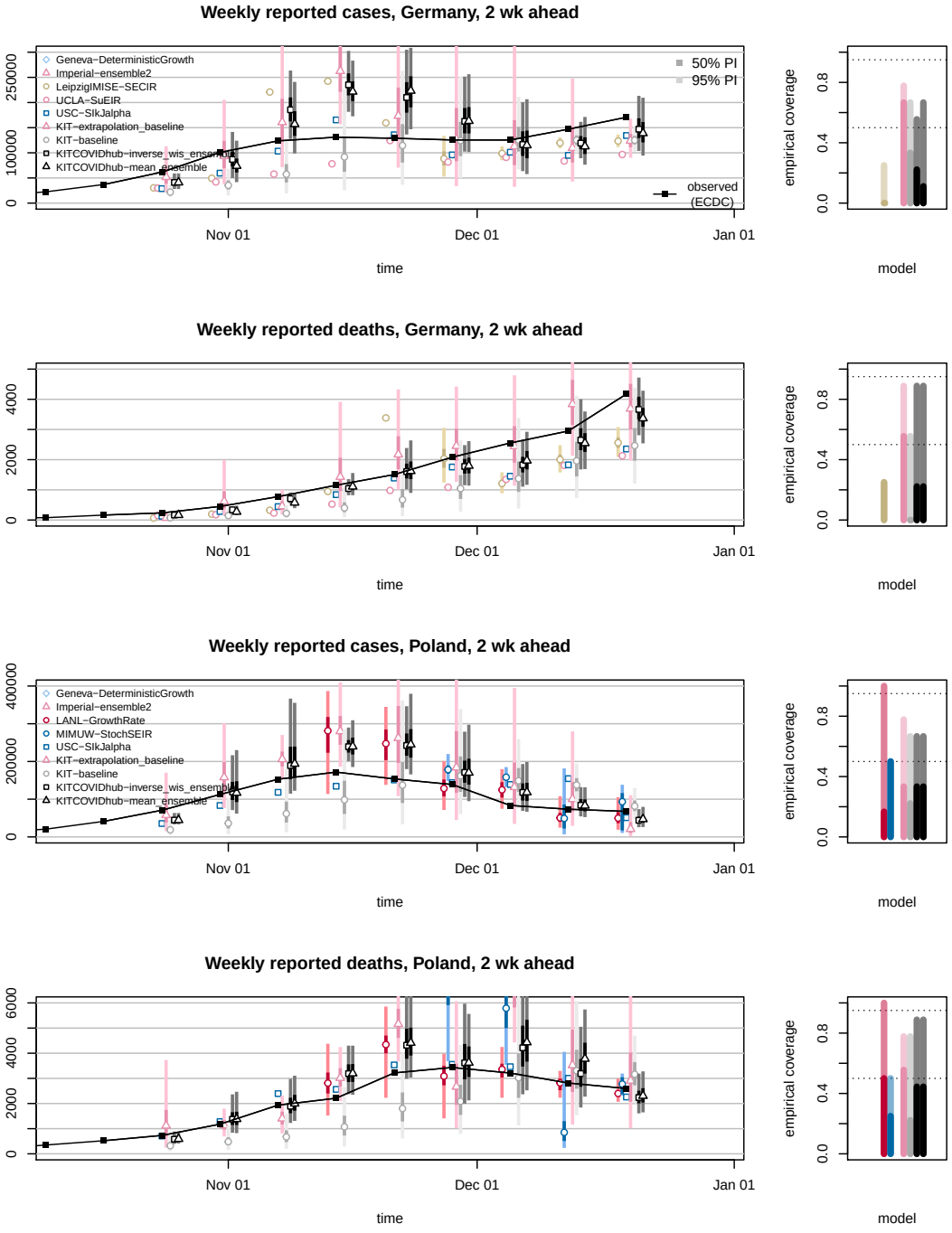

Figure 10: Two-week-ahead forecasts of incident cases and deaths in Germany and Poland (left). Displayed are predictive medians, 50% and 95% prediction intervals for models not shown in Figure 4. Coverage plots (right) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

Table 5: Detailed summary of forecast evaluation for Germany (based on JHU data)

Germany, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	13,351	9,631	3/10	7/10	33,588	25,354	2/9	3/9	13,351	9,631	3/10	7/10	46,506	34,745	2/9	4/9
epiforecasts-EpiNow2	9,879	6,557	6/10	8/10	35,103	26,382	4/9	6/9	9,879	6,557	6/10	8/10	43,106	31,777	5/9	7/9
FIAS_FZJ-Epi1Ger	7,377	4,830	5/10	10/10	29,886	20,831	4/9	7/9	14,775	9,553	3/10	7/10	38,614	27,893	4/9	6/9
Geneva-DeterministicGrowth	14,173								14,173							
ITWW-county_repro	34,825	29,317	0/10	2/10	63,812	52,950	0/9	2/9	34,476	28,969	0/10	2/10	100,873	84,338	0/9	1/9
LANL-GrowthRate	38,679*	23,627*	4/7	7/7	79,788*	44,690*	2/6	6/6	39,734	27,919	4/10	7/10	120,334	82,994	2/9	5/9
LeipzigIMISE-SECIR	20,064		1/5	3/5	51,576		0/4	1/4	36,089	32,015	0/10	1/10	93,828	84,611	1/9	2/9
MIT_CovidAnalytics-DELPHI	40,959*	29,977*	2/8	4/8	82,336*	64,332*	0/7	2/7								
UCLA-SuEIR	28,331				50,970				28,331				76,523			
USC-SIkJalpha	21,935		1/1	1/1	34,150				23,474		1/1	1/1	53,480			
KIT-baseline	21,980	15,297	4/10	8/10	35,913	28,357	3/9	5/9	21,980	15,297	4/10	8/10	53,840	42,904	3/9	5/9
KIT-extrapolation_baseline	12,851	10,585	7/10	10/10	37,194	26,776	5/9	7/9	12,851	10,585	7/10	10/10	47,861	35,388	7/9	7/9
KIT-time_series_baseline	14,910	11,115	6/10	9/10	43,955	28,963	5/9	8/9	14,910	11,115	6/10	9/10	59,721	39,560	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	14,205	9,209	3/10	9/10	42,759	28,516	1/9	6/9	13,525	9,009	4/10	8/10	53,689	35,754	2/9	7/9
KITCOVIDhub-mean_ensemble	17,484	10,994	3/10	8/10	42,910	28,407	2/9	6/9	16,171	10,637	3/10	7/10	57,842	38,266	1/9	6/9
KITCOVIDhub-median_ensemble	12,271	8,179	4/10	9/10	38,316	26,306	3/9	7/9	13,277	9,142	4/10	7/10	50,155	35,157	2/9	6/9

Germany, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	236	159	4/10	7/10	390	273	3/9	6/9	236	159	4/10	7/10	545	359	2/9	7/9
epiforecasts-EpiNow2	159	110	5/10	8/10	357	224	4/9	7/9	159	110	5/10	8/10	485	304	5/9	7/9
FIAS_FZJ-Epi1Ger	307	270	2/10	4/10	579	488	1/9	3/9	244	208	0/10	3/10	779	660	2/9	4/9
Geneva-DeterministicGrowth	409								409							
Imperial-ensemble2	306	228	2/10	5/10					303	226	2/10	5/10				
ITWW-county_repro	420	404	0/10	2/10	594	538	0/9	1/9	419	403	0/10	2/10	924	855	0/9	2/9
LANL-GrowthRate	234*	152*	4/7	7/7	500*	359*	2/6	5/6	218	152	4/10	7/10	658	500	3/9	5/9
LeipzigIMISE-SECIR	653		0/5	1/5	815		1/4	1/4	1,199	1,032	0/10	1/10	2,247	1,883	0/9	1/9
MIT_CovidAnalytics-DELPHI	499*	383*	2/8	3/8	401*	317*	1/7	5/7	473*	358*	2/8	3/8	635*	537*	1/7	3/7
UCLA-SuEIR	508				884				508				1,281			
USC-SIkJalpha	540		0/1	0/1	657				552		0/1	0/1	1,066			
KIT-baseline	531	301	0/10	9/10	892	568	0/9	5/9	531	301	0/10	9/10	1,258	808	0/9	5/9
KIT-extrapolation_baseline	181	137	7/10	9/10	385	250	5/9	8/9	181	137	7/10	9/10	483	326	6/9	8/9
KIT-time_series_baseline	213	180	6/10	9/10	592	410	4/9	8/9	213	180	6/10	9/10	803	558	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	218	145	2/10	8/10	302	183	1/9	8/9	214	139	2/10	8/10	474	287	1/9	6/9
KITCOVIDhub-mean_ensemble	256	171	1/10	9/10	345	212	2/9	8/9	267	181	2/10	9/10	544	333	1/9	8/9
KITCOVIDhub-median_ensemble	251	169	2/10	8/10	381	256	2/9	7/9	242	163	3/10	8/10	543	339	2/9	8/9

Table 6: Detailed summary of forecast evaluation for Poland (based on JHU data)

Poland, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	16,105	11,587	4/10	8/10	41,186	28,453	1/9	5/9	16,128	11,611	4/10	8/10	59,079	39,405	1/9	6/9
epiforecasts-EpiNow2	9,147	6,693	7/10	7/10	36,477	24,601	2/9	6/9	9,170	6,701	7/10	7/10	42,849	28,720	6/9	7/9
Geneva-DeterministicGrowth	10,817								10,888							
ICM-agentModel			2/4	4/4			0/3	1/3			1/4	3/4			0/3	2/3
ITWW-county_repro	19,441	16,339	1/10	4/10	37,152	31,140	2/9	2/9	18,519	15,413	1/10	3/10	57,467	47,412	1/9	3/9
LANL-GrowthRate	13,494*	8,918*	5/7	7/7	48,693*	28,561*	1/6	6/6	15,205	10,361	5/10	8/10	66,280	41,022	2/9	8/9
MIMUW-StochSEIR			3/5	4/5			2/4	2/4			2/5	4/5			1/4	2/4
MIT_CovidAnalytics-DELPHI	30,505*	22,955*	3/9	5/9	61,303*	46,983*	1/8	4/8								
MOCOS-agent1	15,732	11,631	1/10	4/10	32,235	26,356	1/9	3/9	15,732	11,631	1/10	4/10	46,214	36,067	1/9	3/9
USC-SIkJalpha	19,270		0/1	1/1	36,059				22,480		0/1	1/1	57,255			
KIT-baseline	31,605	20,485	5/10	9/10	55,931	38,168	2/9	6/9	31,676	20,521	5/10	9/10	87,597	60,347	2/9	6/9
KIT-extrapolation_baseline	18,333	12,029	7/10	10/10	55,685	34,991	3/9	8/9	18,311	12,025	7/10	10/10	77,269	46,459	3/9	9/9
KIT-time_series_baseline	22,502	14,790	5/10	10/10	60,704	39,562	4/9	9/9	22,480	14,787	5/10	10/10	85,192	54,639	4/9	9/9
KITCOVIDhub-inverse_wis_ensemble	14,191	9,315	5/10	8/10	37,174	26,411	3/9	6/9	14,325	9,066	5/10	8/10	50,096	34,165	2/9	6/9
KITCOVIDhub-mean_ensemble	13,849	9,086	4/10	9/10	37,831	25,362	2/9	6/9	14,511	8,968	5/10	8/10	50,731	32,987	2/9	7/9
KITCOVIDhub-median_ensemble	15,236	9,767	6/10	9/10	40,453	26,329	2/9	6/9	16,541	10,373	4/10	8/10	55,827	35,166	1/9	5/9

Poland, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	303	191	3/10	9/10	625	408	2/9	7/9	303	191	3/10	9/10	911	585	2/9	8/9
epiforecasts-EpiNow2	343	244	5/10	7/10	1,066	771	3/9	5/9	344	244	5/10	7/10	1,439	1,021	3/9	5/9
Geneva-DeterministicGrowth	180								179							
ICM-agentModel	808*	731*	1/8	2/8	1,921*	1,298*	0/7	2/7	1,234*	788*	0/8	3/8	3,000*	1,949*	0/7	4/7
Imperial-ensemble2	379	235	6/10	7/10					351	207	6/10	7/10				
ITWW-county_repro	465	429	1/10	1/10	652	564	0/9	2/9	458	423	2/10	2/10	1,099	973	0/9	3/9
LANL-GrowthRate	236*	161*	4/7	7/7	383*	241*	4/6	6/6	236	158	4/10	9/10	655	430	3/9	8/9
MIMUW-StochSEIR			2/5	4/5			1/4	2/4			2/5	4/5			0/4	4/4
MIT_CovidAnalytics-DELPHI	467*	306*	2/9	7/9	621*	417*	1/8	6/8	552*	393*	2/9	7/9	990*	717*	1/8	4/8
MOCOS-agent1	205	152	8/10	10/10	393	257	5/9	8/9	205	152	8/10	10/10	533	362	7/9	9/9
USC-SIkJalpha	202		0/1	1/1	212				252		0/1	0/1	283			
KIT-baseline	504	313	5/10	10/10	882	591	2/9	6/9	503	313	5/10	10/10	1,365	915	2/9	5/9
KIT-extrapolation_baseline	412	280	6/10	8/10	995	712	5/9	7/9	411	280	6/10	8/10	1,422	993	4/9	7/9
KIT-time_series_baseline	528	341	8/10	10/10	1,343	874	5/9	8/9	529	341	8/10	10/10	1,909	1,235	5/9	8/9
KITCOVIDhub-inverse_wis_ensemble	207	141	7/10	10/10	476	307	4/9	9/9	231	154	7/10	10/10	683	443	4/9	8/9
KITCOVIDhub-mean_ensemble	227	150	7/10	10/10	558	358	4/9	9/9	250	162	6/10	10/10	793	512	4/9	9/9
KITCOVIDhub-median_ensemble	195	137	6/10	10/10	460	285	4/9	9/9	215	150	6/10	10/10	686	443	4/9	8/9

Table 7: Summary of forecast evaluation for ensembles without plausibility checks of members (based on ECDC data)

Germany, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
KITCOVIDhub-inverse_wis_ensemble_all	13,431	9,026	4/10	9/10	39,275	25,413	1/9	6/9	28,345	19,130	1/10	7/10	55,290	36,153	2/9	7/9
KITCOVIDhub-mean_ensemble_all	15,554	10,086	4/10	9/10	40,120	25,588	1/9	6/9	16,068	10,633	4/10	7/10	54,550	35,351	2/9	6/9
KITCOVIDhub-median_ensemble_all	11,240	8,096	6/10	9/10	36,823	24,379	3/9	7/9	13,511	9,756	6/10	7/10	51,242	35,713	2/9	6/9

Germany, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
KITCOVIDhub-inverse_wis_ensemble_all	177	111	6/10	9/10	234	148	4/9	8/9	845	673	0/10	8/10	1,065	756	1/9	8/9
KITCOVIDhub-mean_ensemble_all	183	126	6/10	8/10	263	166	4/9	8/9	236	160	4/10	6/10	472	298	3/9	6/9
KITCOVIDhub-median_ensemble_all	185	132	6/10	8/10	332	221	3/9	7/9	196	134	4/10	7/10	434	277	3/9	7/9

Poland, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
KITCOVIDhub-inverse_wis_ensemble_all	12,100	8,233	4/10	9/10	36,692	23,618	3/9	7/9	11,951	7,494	5/10	10/10	44,256	28,726	2/9	8/9
KITCOVIDhub-mean_ensemble_all	11,788	8,011	4/10	9/10	37,031	23,152	2/9	8/9	11,649	7,266	5/10	10/10	43,910	28,304	3/9	9/9
KITCOVIDhub-median_ensemble_all	13,597	8,839	5/10	9/10	39,156	24,368	2/9	7/9	13,365	8,622	5/10	9/10	48,278	29,298	3/9	7/9

Poland, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
KITCOVIDhub-inverse_wis_ensemble_all	188	143	6/10	10/10	467	302	5/9	9/9	384	247	2/10	9/10	656	417	5/9	9/9
KITCOVIDhub-mean_ensemble_all	204	150	7/10	9/10	593	363	4/9	9/9	220	157	7/10	9/10	802	498	3/9	9/9
KITCOVIDhub-median_ensemble_all	202	141	6/10	10/10	428	278	6/9	9/9	194	138	8/10	10/10	620	413	6/9	9/9

F Results for three- and four-week-ahead forecasts

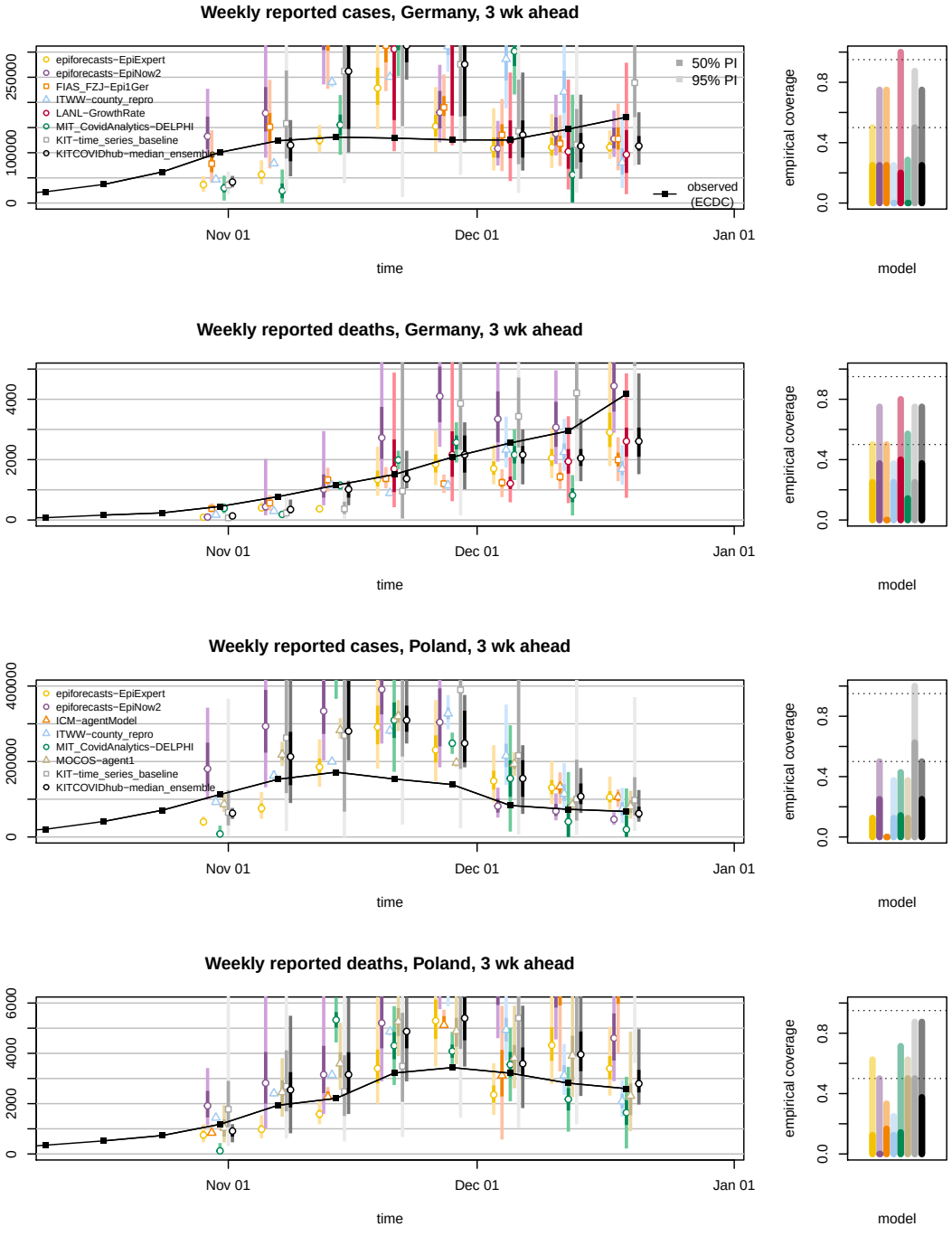

Figure 11: Three-week-ahead forecasts of incident cases and deaths in Germany and Poland (left). Displayed are predictive medians, 50% and 95% prediction intervals. Coverage plots (right) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

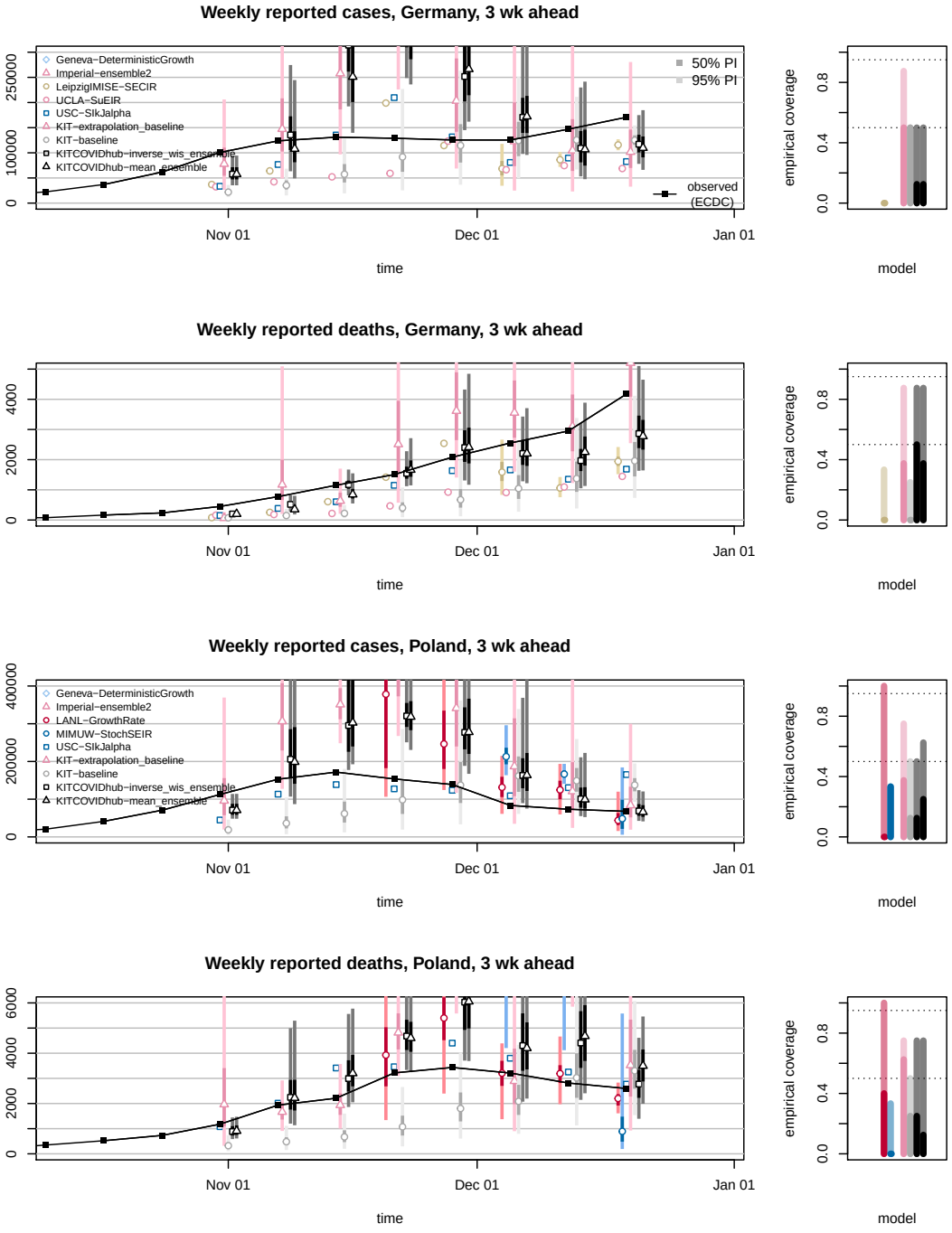

Figure 12: Three-week-ahead forecasts of incident cases and deaths in Germany and Poland (left). Displayed are predictive medians, 50% and 95% prediction intervals for models not shown in Figure 11. Coverage plots (right) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

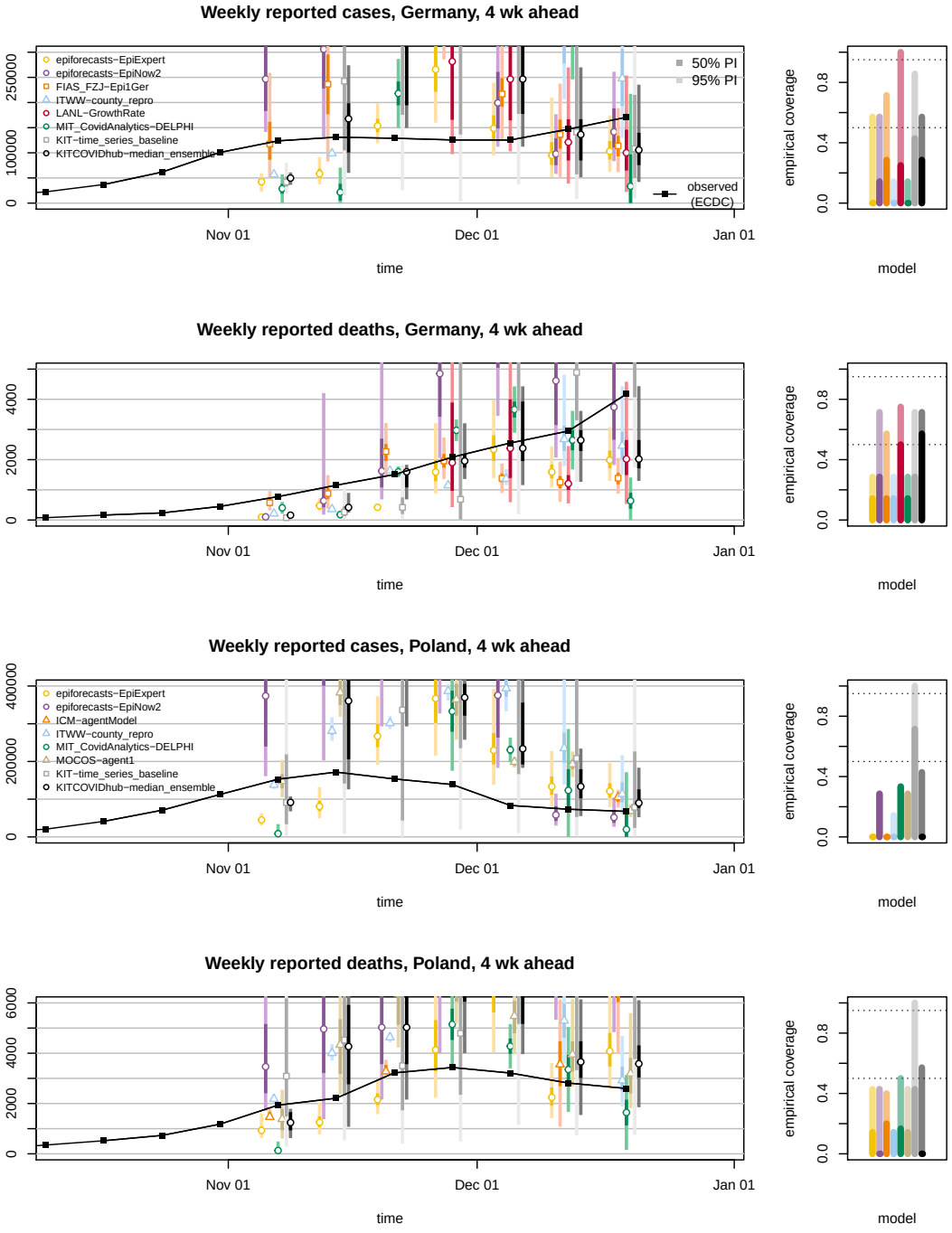

Figure 13: Four-week-ahead forecasts of incident cases and deaths in Germany and Poland (left). Displayed are predictive medians, 50% and 95% prediction intervals. Coverage plots (right) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

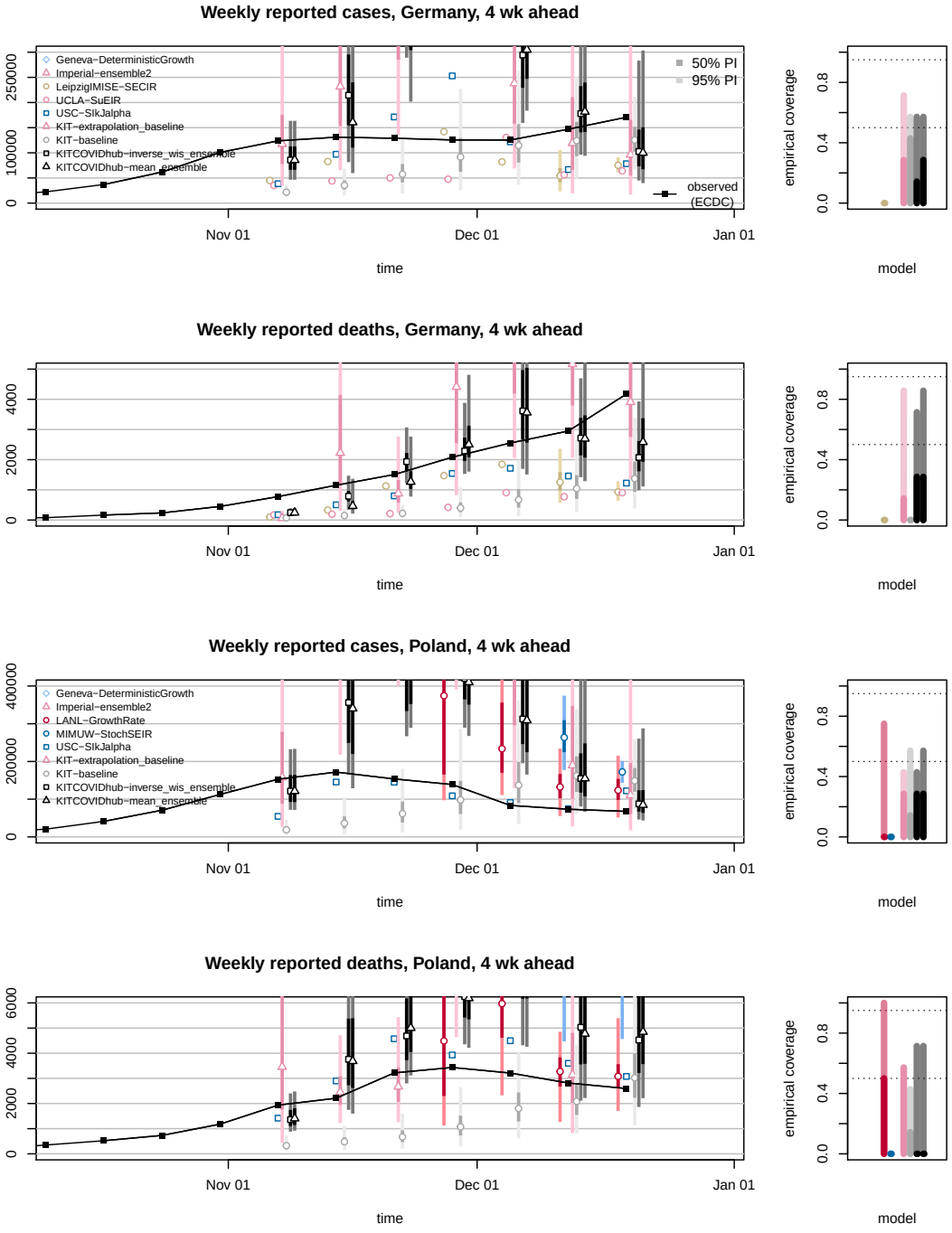

Figure 14: Four-week-ahead forecasts of incident cases and deaths in Germany and Poland (left). Displayed are predictive medians, 50% and 95% prediction intervals for models not shown in Figure 13. Coverage plots (right) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

Table 8: Detailed summary of forecast evaluation for Germany, 3 and 4 weeks ahead (based on ECDC data)

Germany, cases

Model	3 wk ahead inc				4 wk ahead inc				3 wk ahead cum				4 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	47,162	35,223	2/8	4/8	65,906	48,401	0/7	4/7	85,859	61,545	2/8	4/8	143,536	102,969	2/7	4/7
epiforecasts-EpiNow2	98,948	74,822	2/8	6/8	235,461	178,179	1/7	4/7	148,866	111,538	3/8	6/8	394,730	297,702	1/7	5/7
FIAS_FZJ-Epi1Ger	74,550	55,218	2/8	6/8	149,451	114,912	2/7	5/7	116,086	86,018	4/8	6/8	271,463	205,908	2/7	4/7
Geneva-DeterministicGrowth																
ITWW-county_repro	105,132	89,201	0/8	2/8	147,570	129,984	0/7	1/7	214,867	181,586	0/8	1/8	369,098	321,879	0/7	1/7
LANL-GrowthRate			1/5	5/5			1/4	4/4	218,206	154,962	2/8	5/8	328,115	250,254	2/7	4/7
LeipzigIMISE-SECIR	82,104		0/3	0/3	140,275		0/2	0/2	171,325	154,467	1/8	2/8	296,564	268,332	2/7	2/7
MIT_CovidAnalytics-DELPHI	139,142*	115,333*	0/7	2/7	210,370	179,364	0/7	1/7								
UCLA-SuEIR	66,768				76,415				132,374				198,546			
USC-SIkJalpha	49,446				66,436				96,902				140,716			
KIT-baseline	44,706	36,259	4/8	4/8	54,563	42,708	3/7	4/7	85,838	71,384	3/8	4/8	136,280	116,029	2/7	4/7
KIT-extrapolation_baseline	82,243	57,464	4/8	7/8	165,710	117,658	2/7	5/7	125,152	90,266	4/8	7/8	291,137	210,908	4/7	6/7
KIT-time_series_baseline	91,848	64,668	4/8	7/8	162,293	134,547	3/7	6/7	154,663	105,208	3/8	7/8	320,671	242,968	3/7	6/7
KITCOVIDhub-inverse_wis_ensemble	90,487	66,261	1/8	4/8	171,254	128,701	1/7	4/7	136,583	95,759	3/8	5/8	278,362	209,815	3/7	5/7
KITCOVIDhub-mean_ensemble	82,294	58,349	1/8	4/8	138,225	105,500	2/7	4/7	137,562	94,440	3/8	5/8	262,811	193,455	3/7	4/7
KITCOVIDhub-median_ensemble	79,238	54,347	2/8	6/8	129,400	94,232	2/7	4/7	125,538	90,852	2/8	5/8	252,946	194,526	3/7	4/7

Germany, deaths

Model	3 wk ahead inc				4 wk ahead inc				3 wk ahead cum				4 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	615	435	2/8	4/8	955	791	1/7	2/7	981	669	2/8	5/8	1,850	1,399	2/7	3/7
epiforecasts-EpiNow2	655	465	3/8	6/8	1,511	1,049	2/7	5/7	1,044	723	3/8	6/8	2,532	1,768	2/7	5/7
FIAS_FZJ-Epi1Ger	811	689	0/8	4/8	1,003	828	1/7	4/7	1,467	1,258	1/8	4/8	2,153	1,800	0/7	4/7
Geneva-DeterministicGrowth																
Imperial-ensemble2																
ITWW-county_repro	713	618	2/8	3/8	798	676	1/7	2/7	1,345	1,181	2/8	2/8	1,740	1,498	1/7	3/7
LANL-GrowthRate			2/5	4/5			2/4	3/4	1,364	1,109	2/8	4/8	2,238	2,015	2/7	3/7
LeipzigIMISE-SECIR	883		0/3	1/3	1,162		0/2	0/2	2,722	2,301	1/8	2/8	2,780	2,404	1/7	2/7
MIT_CovidAnalytics-DELPHI	593*	480*	1/7	4/7	1,039	899	1/7	2/7	1,211*	1,061*	0/7	1/7	2,201	2,026	0/7	0/7
UCLA-SuEIR	1,279				1,659				2,273				3,615			
USC-SIkJalpha	877				1,110				1,635				2,344			
KIT-baseline	1,218	871	0/8	2/8	1,608	1,291	0/7	0/7	2,186	1,609	0/8	0/8	3,538	2,972	0/7	0/7
KIT-extrapolation_baseline	752	492	3/8	7/8	1,526	1,014	1/7	6/7	1,252	818	3/8	7/8	2,666	1,770	1/7	6/7
KIT-time_series_baseline	1,092	802	2/8	6/8	1,712	1,738	2/7	5/7	1,727	1,278	3/8	7/8	2,934	2,807	2/7	5/7
KITCOVIDhub-inverse_wis_ensemble	440	281	4/8	7/8	704	481	2/7	5/7	810	512	2/8	6/8	1,259	897	4/7	4/7
KITCOVIDhub-mean_ensemble	488	305	3/8	7/8	676	443	2/7	6/7	880	571	3/8	6/8	1,301	856	3/7	5/7
KITCOVIDhub-median_ensemble	493	341	3/8	6/8	599	451	4/7	5/7	897	603	3/8	6/8	1,310	992	3/7	5/7

Table 9: Detailed summary of forecast evaluation for Poland, 3 and 4 weeks ahead (based on ECDC data)

Poland, cases

Model	3 wk ahead inc				4 wk ahead inc				3 wk ahead cum				4 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	69,036	50,972	1/8	1/8	114,361	89,227	0/7	0/7	123,420	85,039	1/8	5/8	236,085	166,137	0/7	2/7
epiforecasts-EpiNow2	100,091	71,090	2/8	4/8	257,145	186,780	2/7	2/7	146,121	98,515	3/8	6/8	422,386	296,595	2/7	4/7
Geneva-DeterministicGrowth																
ICM-agentModel			0/2	0/2			0/1	0/1			0/2	1/2			0/1	0/1
ITWW-county_repro	69,096	60,024	1/8	3/8	148,272	133,453	0/7	1/7	115,872	96,816	1/8	4/8	271,133	235,131	0/7	1/7
LANL-GrowthRate			0/5	5/5			0/4	3/4	147,981	88,161	1/8	7/8	278,420	168,038	1/7	5/7
MIMUW-StochSEIR			1/3	1/3			0/2	0/2			1/3	1/3			0/2	0/2
MIT_CovidAnalytics-DELPHI	113,802*	89,856*	1/7	3/7	177,706*	151,123*	2/6	2/6								
MOCOS-agent1	70,362	61,332	1/8	3/8	140,691	126,585	2/7	2/7	116,589	95,610	0/8	2/8	263,653	228,313	0/7	2/7
USC-SIkJalpha	45,309				32,410				99,459				110,078			
KIT-baseline	75,183	55,427	1/8	4/8	89,395	65,034	1/7	4/7	159,350	119,908	1/8	3/8	238,036	185,642	1/7	3/7
KIT-extrapolation_baseline	125,846	89,406	3/8	6/8	278,666	212,898	2/7	3/7	206,056	138,276	3/8	6/8	505,952	366,502	2/7	4/7
KIT-time_series_baseline	123,714	91,450	5/8	8/8	228,443	216,494	5/7	7/7	217,544	148,705	5/8	8/8	476,907	380,524	5/7	7/7
KITCOVIDhub-inverse_wis_ensemble	78,949	58,408	1/8	4/8	160,008	130,808	2/7	3/7	115,211	84,219	2/8	5/8	266,961	199,966	1/7	5/7
KITCOVIDhub-mean_ensemble	78,772	57,123	2/8	5/8	156,341	126,412	2/7	4/7	115,204	81,636	2/8	6/8	265,775	194,441	1/7	5/7
KITCOVIDhub-median_ensemble	74,351	58,336	2/8	4/8	144,957	116,880	0/7	3/7	126,339	79,697	1/8	5/8	275,236	175,685	1/7	4/7

Poland, deaths

Model	3 wk ahead inc				4 wk ahead inc				3 wk ahead cum				4 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	901	605	1/8	5/8	1,329	923	1/7	3/7	1,828	1,164	1/8	6/8	3,171	2,082	1/7	4/7
epiforecasts-EpiNow2	2,642	1,873	0/8	4/8	5,317	3,822	0/7	3/7	4,316	3,033	1/8	4/8	9,641	6,844	0/7	4/7
Geneva-DeterministicGrowth																
ICM-agentModel	1,854*	1,455*	1/6	2/6	1,850*	1,520*	1/5	2/5	4,376*	2,912*	1/6	2/6	4,584*	3,329*	1/5	3/5
Imperial-ensemble2																
ITWW-county_repro	1,113	964	1/8	2/8	2,016	1,808	1/7	1/7	2,375	2,092	0/8	2/8	4,592	4,123	0/7	1/7
LANL-GrowthRate			2/5	5/5			2/4	4/4	1,517	962	4/8	7/8	2,984	1,999	3/7	6/7
MIMUW-StochSEIR			0/3	1/3			0/2	0/2			0/3	1/3			0/2	0/2
MIT_CovidAnalytics-DELPHI	1,122*	843*	1/7	5/7	1,851*	1,503*	1/6	3/6	2,076*	1,742*	1/7	1/7	3,899*	3,501*	0/6	1/6
MOCOS-agent1	940	595	4/8	5/8	1,883	1,279	1/7	3/7	1,450	904	5/8	7/8	3,497	2,114	3/7	4/7
USC-SIkJalpha	474				801				597				1,445			
KIT-baseline	1,208	876	2/8	4/8	1,544	1,215	1/7	3/7	2,368	1,793	1/8	4/8	3,986	3,258	1/7	2/7
KIT-extrapolation_baseline	1,995	1,534	5/8	6/8	4,093	3,270	4/7	4/7	3,537	2,602	4/8	6/8	7,690	6,040	3/7	4/7
KIT-time_series_baseline	2,235	1,518	4/8	7/8	3,084	2,214	3/7	7/7	3,799	2,559	5/8	7/8	5,787	3,951	4/7	7/7
KITCOVIDhub-inverse_wis_ensemble	1,039	665	2/8	6/8	2,205	1,479	0/7	5/7	1,668	1,094	4/8	8/8	3,850	2,525	2/7	5/7
KITCOVIDhub-mean_ensemble	1,165	729	1/8	6/8	2,232	1,493	0/7	5/7	1,978	1,272	3/8	6/8	4,109	2,713	1/7	5/7
KITCOVIDhub-median_ensemble	895	547	3/8	7/8	1,871	1,190	0/7	4/7	1,645	1,055	4/8	7/8	3,656	2,259	2/7	4/7