National and subnational short-term forecasting of COVID-19 in Germany and Poland during early 2021

Johannes Bracher[1,2,*], Daniel Wolffram[1,2,3], Jannik Deuschel[1], Konstantin Görgen[1], Jakob L. Ketterer[1], Alexander Ullrich[4], Sam Abbott[5], Maria V. Barbarossa[6], Dimitris Bertsimas[7], Sangeeta Bhatia[8], Marcin Bodych[9], Nikos I. Bosse[5], Jan Pablo Burgard[10] Lauren Castro[11] Geoffrey Fairchild[11] Jochen Fiedler[12], Jan Fuhrmann[13] Sebastian Funk[5], Anna Gambin[14] Krzysztof Gogolewski[14] Stefan Heyder[15], Thomas Hotz[15] Yuri Kheifetz[16] Holger Kirsten[16] Tyll Krueger[9], Elena Krymova[17], Neele Leithäuser[12] Michael L. Li[18] Jan H. Meinke[19] Błażej Miasojedow[14] Isaac J. Michaud[20], Jan Mohring[12] Pierre Nouvellet[21] Jedrzej M. Nowosielski[22] Tomasz Ozanski[9], Maciej Radwan[22] Franciszek Rakowski[22] Markus Scholz[16] Saksham Soni[18], Ajitesh Srivastava[23] Tilmann Gneiting[2,24] Melanie Schienle[1,*]

[*] 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), Heidelberg, Germany
[3]HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[4]Robert Koch Institute (RKI), Berlin, Germany
[5]London School of Hygiene and Tropical Medicine, London, UK
[6]Frankfurt Institute for Advanced Studies, Frankfurt, Germany
[7]Sloan School of Management, Massachusetts Institute of Technology, Cambridge MA, USA
[8]MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), Imperial College London, London, UK
[9]Wroclaw University of Science and Technology, Wroclaw, Poland
[10]Economic and Social Statistics Department, University of Trier, Trier, Germany
[11]Information Systems and Modeling, Los Alamos National Laboratory, Los Alamos, USA
[12]Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany
[13]Institute for Applied Mathematics, University of Heidelberg, Heidelberg, Germany
[14]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland
[15]Institute of Mathematics, Technische Universität Ilmenau, Ilmenau, Germany
[16]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany
[17]Swiss Data Science Center, ETH Zürich and EPF Lausanne, Zürich, Switzerland
[18]Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA
[19]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany
[20]Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, USA
[21]School of Life Sciences, University of Sussex, Brighton, UK
[22]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland
[23]Ming Hsieh Department of Computer and Electrical Engineering, University of Southern California, Los Angeles, USA
[24]Institute for Stochastics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Abstract

Background. During the COVID-19 pandemic there has been a strong interest in forecasts of the short-term development of epidemiological indicators to inform decision makers. In this study we evaluate probabilistic real-time predictions of confirmed cases and deaths from COVID-19 in Germany and Poland for the period from January through April 2021.

Methods. We evaluate probabilistic real-time predictions of confirmed cases and deaths from COVID-19 in Germany and Poland. These were issued by 15 different forecasting models, run by independent research teams. Moreover, we study the performance of combined ensemble forecasts. Evaluation of probabilistic forecasts is based on proper scoring rules, along with interval coverage proportions to assess forecast calibration. The presented work is part of a pre-registered evaluation study and covers the period from January through April 2021.

Results. We find that many, though not all, models outperform a simple baseline model up to four weeks ahead for the considered targets. Ensemble methods (i.e., combinations of different available forecasts) show very good relative performance. The addressed time period is characterized by rather stable non-pharmaceutical interventions in both countries, making short-term predictions more straightforward than in previous periods. However, major trend changes in reported cases, like the rebound in cases due to the rise of the B.1.1.7 (alpha) variant in March 2021, prove challenging to predict.

Conclusions. Multi-model approaches can help to improve the performance of epidemiological forecasts. However, while death numbers can be predicted with some success based on current case and hospitalization data, predictability of case numbers remains low beyond quite short time horizons. Additional data sources including sequencing and mobility data, which were not extensively used in the present study, may help to improve performance.

Plain language summary

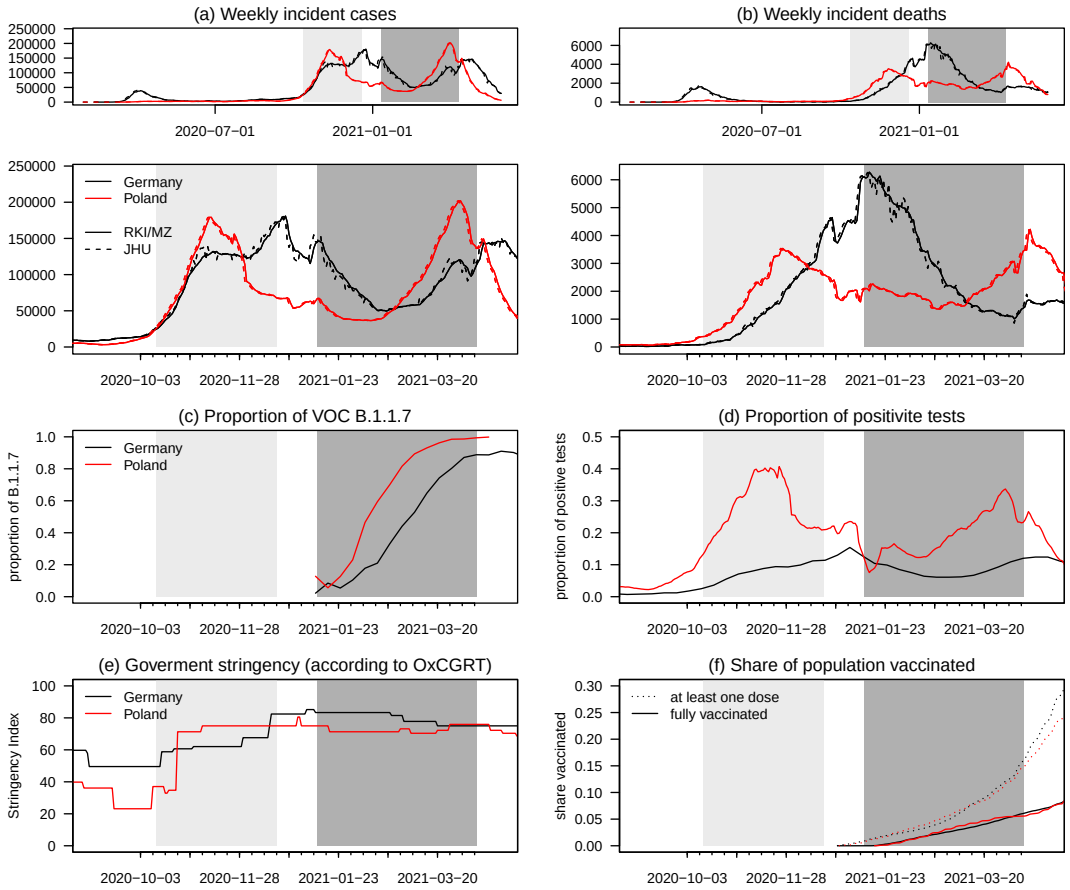
The goal of this study is to assess the quality of forecasts of weekly case and death numbers of COVID-19 in Germany and Poland during the period of January through April 2021. We focus on real-time forecasts at time horizons of one and two weeks ahead created by fourteen independent teams. Forecasts are systematically evaluated taking uncertainty ranges of predictions into account. We find that combining different forecasts into ensembles can improve the quality of predictions, but especially case numbers proved very challenging to predict beyond quite short time windows. Additional data sources, in particular genetic sequencing data, may help to improve forecasts in the future.

Introduction

Short-term forecasts of infectious diseases and longer-term scenario projections provide complementary perspectives to inform public health decision making. Both have received considerable attention during the COVID-19 pandemic and are increasingly embraced by public health agencies. This is illustrated by the US COVID-19 Forecast[1,2] and Scenario Modelling Hubs,[3] supported by the US Centers for Disease Control and Prevention, as well as the more recent European COVID-19 Forecast Hub,[4] supported by the European Center for Disease Prevention and Control (ECDC). The Forecast Hub concept, building on pre-pandemic collaborative disease forecasting projects like FluSight,[5] the DARPA Chikungunya Challenge[6] or the Dengue Forecasting Project[7] aims to provide a broad picture of existing short-term projections in real time, making the agreement or disagreement between different models visible. Also, it forms the basis for a systematic evaluation of performance. This is a prerequisite for model consolidation and improvement, and a need repeatedly expressed.[8] It has been highlighted that such modelling studies should be prospective[9] and ideally follow pre-registered protocols[10] in order to prevent selective reporting and hindsight bias (i.e., the tendency to overstate the predictability of past events in hindsight).

We here report on the second part of a prospective disease forecasting study, pre-registered on 8 October 2020[11] and including forecasts made between 11 January 2021 and 29 March 2021 (with last observed values running through April; twelve weeks of forecasting). It is based on the German and Polish COVID-19 Forecast Hub (https://kitmetricslab.github.io/forecasthub/), which gathers and stores forecasts in real time. This platform was launched in close exchange with the US COVID-19 Forecast Hub in June 2020. In April 2021 it was largely merged into the European COVID-19 Forecast Hub, shortly after the latter had been initiated by ECDC. During our study period, fifteen independent modelling teams provided forecasts of cases and deaths by appearance in publicly available national-level data, provided either by national

Figure 1: Overview of relevant epidemiological time series.



Reported cases (a) and deaths (b) in Germany (black) and Poland (red) according to Robert Koch Institute, the Polish Ministry of Health (MZ; solid lines) and Johns Hopkins CSSE (dashed). Additional panels show (c) the share of cases due to the B.1.1.7 (Alpha) variant, (d) the proportion of all performed PCR tests which turned out positive, (e) the overall level of non-pharmaceutical interventions as measured by the Oxford Coronavirus Government Response Tracker (OxCGRT) Stringency Index, and (f) the population shares having received at least one vaccination dose (dotted) and complete vaccination (solid). The dark grey area indicates the period addressed in the present manuscript, the light grey area the one from Bracher et al.[16]

health authorities (Robert Koch Institute, RKI[12] or the Polish Ministry of Health, MZ;[13] the primary data source) or the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE;[14] and[15]). As specified in our study protocol, we report results on forecasts up to a horizon of four weeks, but focus on forecasts one and two weeks ahead. While we acknowledge the relevance of longer horizons for planning purposes, we argue that factors like changing non-pharmaceutical interventions and emergence of new variants limit meaningful forecasts (as opposed to scenarios) to rather short time horizons, especially for cases. Also, we focus almost exclusively on incident quantities, as their cumulative counterparts have almost completely vanished from any public discussion.

The time series of cases and deaths in both countries are displayed in panels (a) and (b) of Figure 1. The study period covered in this paper is marked in dark grey, while the light grey area represents the time span addressed in the first part of our study.[16] Our study period contains the transition from the original wild type variant of the virus to the B.1.1.7 variant (later called Alpha). Panel (c) of Figure 1 shows the estimated weekly percentages of all cases which were due to the B.1.1.7 variant in Germany[17] and Poland[18,19] in calendar weeks 4–12. Panel (d) shows the proportion of all performed PCR tests which turned out positive. While in Germany the curve follows a U-shape similar to the case incidence curve, the test

positivity rate continuously increased in Poland, peaking at 33%. Panel (e) shows the Oxford Coronavirus Government Response Tracker (OxCGRT) Stringency Index.[20] It can be seen that compared to the first part of our study, the level of non-pharmaceutical interventions was rather stable at a high level during the second period. We note, however, that on 27 March a new set of restrictions was added in Poland (closure of daycare centers, hair salons and sports facilities, among others), which is not reflected very strongly in the stringency index. The start of vaccination rollout in both countries coincides with the start of our study period. However, by its end only roughly one sixth of the population of both countries had received a first dose, and roughly one twentieth had received two doses (with the role of the one-dose Johnson and Johnson vaccine negligible in both countries); see panel (f).

We find that averaged over the second evaluation period, most though not all of the compared models were able to outperform a naïve baseline model. Heterogeneity between forecasts from different models was considerable. Ensemble forecasts combining different available predictions achieved very good performance relative to single-model forecasts. However, most models, including the ensemble, did not anticipate changes in trend well, in particular for cases. Pooling results over both evaluation periods we find that ensemble forecasts for deaths were well-calibrated (i.e., prediction intervals contained the true value roughly as often as intended) even at longer prediction horizons and clearly outperformed baseline and individual models, while for cases this was only the case for one- and to a lesser degree two-week-ahead forecasts.

Methods

The methods described in the following are largely identical to those in the first part[16] of our study, but are presented to ensure self-containedness of the present work.

Targets and submission system. Teams submitted forecasts for weekly incident and cumulative confirmed cases and deaths from COVID-19 via a dedicated public GitHub repository (https://github.com/KITmetricslab/covid19-forecast-hub-de). For certain teams running public dashboards, software scripts were put in place to transfer forecasts to the Forecast Hub repository. Weeks were defined to run from Sunday through Saturday. Each week, teams were asked to submit forecasts using data available up to Monday, with submission possible until Tuesday 3pm Berlin/Warsaw time (the first two daily observations were thus already available at the time of forecasting). Forecasts could either refer to the time series provided by JHU CSSE or those from Robert Koch Institute and the Polish Ministry of Health. All data streams were aggregated by time of appearance in national data, see also Supplementary Note 4 of Bracher et al.[16] Submissions consisted of a point forecast and 23 predictive quantiles (1%, 2.5%, 5%, 10%, ..., 95%, 97.5%, 0.99) for the incident and cumulative weekly quantities. As in previous work[16] we focus on the targets on the incidence scale. These are easier to compare across the different data sources than cumulative numbers which sometimes show systematic shifts.

Evaluation metrics. As forecasts were reported in the form of 11 nested central prediction intervals (plus the predictive median), a natural choice for evaluation is the interval score.[21] For a central prediction interval $[l, u]$ at the level $(1-\alpha)$, thus reaching from the $\alpha/2$ to the $1-\alpha/2$ quantile, it is defined as

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

where χ is the indicator function and y is the realized value. Here, the first term characterizes the spread of the forecast distribution, the second penalizes overprediction (observations fall below the prediction interval) and the third term penalizes underprediction. To assess the full predictive distribution we use the weighted interval score (WIS;[22]). The WIS is a weighted average of interval scores at different nominal levels and the absolute error. For N nested prediction intervals it is defined as

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

where m is the predictive median and in our setting $N = 11$. The WIS is a well-known approximation of the continuous ranked probability score (CRPS;[21]) and generalizes the absolute error to probabilistic forecasts.

Table 1: Forecast models contributed by independent external research teams.

Category	Model	NPI	Test	Variants	Age	DE	PL	Regional	Truth	Pr
Agent-based	ICM-agentModel[48]	✓	✓	✓	✓		✓		MZ	✓
	MOCOS-agent1[49]	✓	✓	✓	✓		✓		JHU	✓
Compartment	CovidAnalytics-DELPHI[50]	✓				✓	✓		JHU	✓
	FIAS_FZJ-Epi1Ger[51]					✓		✓	RKI	✓
	itwm-dSEIR				✓	✓			RKI	✓
	Karlen-pypm[52]			✓		✓		✓	RKI	✓
	LeipzigIMISE-SECIR[53]	✓	✓	✓		✓		(✓)	RKI	✓
	MIMUW-StochSEIR						✓		JHU	✓
	USC-SIkJalpha[54]					✓	✓	✓	RKI/MZ	✓
Growth rate/ renewal eq.	epiforecasts-EpiNow2[55]					✓	✓	✓	RKI/MZ	✓
	SDSC_ISG-TrendModel[56]					✓	✓		JHU	
	ITWW-county_repro[57]				✓	✓	✓	✓	RKI/MZ	✓
	LANL-GrowthRate[58]					✓	✓		JHU	✓
Human judgement	epiforecasts-EpiExpert[59]	(✓)	(✓)	(✓)	(✓)	✓	✓		RKI/MZ	✓
Forecast ensemble	Imperial-ensemble2[60]					✓	✓		RKI	✓

Abbreviations: NPI: Does the forecast model explicitly account for non-pharmaceutical interventions? Test: Does the model account for changing testing strategies? Variants: Does the model accommodate multiple variants? Age: Is the model age-structured? DE, PL: Are forecasts issued for Germany and Poland, respectively? Regional: Were regional-level forecasts for at least one country submitted? Truth: Which truth data source does the model use? Pr: Are forecasts probabilistic (23 quantiles)? Detailed descriptions of the different models can be found in Bracher et al,[16] Supplementary Note 3 and in Supplement S1 of this article.

Its values can be interpreted on the natural scale of the data and measure how far the observed value y is from the predictive distribution (lower values are thus better). For deterministic one-point forecasts the WIS reduces to the absolute error. A useful property of the WIS is that it inherits the decomposition of the interval score into forecast spread, overprediction and underprediction, which makes average scores more interpretable. As secondary measures of forecast quality we use the absolute error to assess the central tendency of forecasts and interval coverage rates of 50% and 95% prediction intervals to assess calibration.

As specified in our study protocol, whenever forecasts from a model were missing for a given week, we imputed the score with the worst (largest) value achieved by any other model for the respective week and target. However, almost all teams provided complete sets of forecasts and very few scores needed imputation.

Submitted models and baselines. During the evaluation period, forecasts from fifteen different models run by fourteen independent teams of researchers were collected. Thirteen of these were already available during the first part of our study, see Table 3 and Supplementary Note 3 of Bracher et al[16] for detailed descriptions. Table 1 provides a slightly extended summary of model properties, including the two new models, itwm-dSEIR and Karlen-pypm; a more detailed description of the latter can be found in Supplement S1.

During the evaluation period, only the ICM-agentModel explicitly accounted for vaccinations (given the low realized vaccination coverage by the end of the study period this aspect likely had limited impact). Only four models (ICM-agentModel, Karlen-pypm, LeipzigIMISE-SECIR and MOCOS-agent1, all only for certain weeks) explicitly accounted for the presence of multiple variants. In contrast to other related projects,[2] none of the models used mobility data or social media data.

To put the results achieved by the submitted models into perspective, the Forecast Hub team generated forecasts from three simple reference models (see also Bracher et al,[16] Supplementary Note 2):

- KIT-baseline: This is a simple last-observation-carried-forward model, i.e., it predicts the last observed value indefinitely into the future. Predictive quantiles are obtained by assuming a negative binomial observation model with a dispersion parameter estimated via maximum likelihood from five recent observations

- KIT-extrapolation_baseline: This model extrapolates exponential growth or decrease if the last three observations are monotonically increasing or decreasing, with a weekly growth rate equal to the one observed between the last and second to last week; if the last three observations are not ordered,

it predicts a plateau. Predictive quantiles are again obtained using a negative binomial observation model and five recent observations.

- KIT-time_series_baseline: This is an exponential smoothing time series model with multiplicative errors as used by Petropoulos and Makridakis[23] to predict COVID-19 cases and deaths. It is implemented using the R package forecast, version 8.12.[24]

As a further external comparison we added publicly available death forecasts by the Institute for Health Metrics and Evaluation (IHME, University of Washington;[25] available under the CC BY-NC 4.0 license). Here, we always used the most recent prediction available on a given forecast date.

Forecast ensembles. The Forecast Hub team used the submitted forecasts to generate three different ensemble forecasts:

- KITCOVIDhub-median_ensemble The α-quantile of the ensemble forecast is obtained as the median of the α-quantiles of the member forecasts.

- KITCOVIDhub-mean_ensemble The α-quantile of the ensemble forecast is obtained as the mean of the α-quantiles of the member forecasts.

- KITCOVIDhub-inverse_wis_ensemble The α-quantile of the ensemble forecast is a convex combination of the α-quantiles of the member forecasts. The weights are chosen inversely proportional to the mean WIS value obtained by the member models over the last six evaluated forecasts (last three one-week-ahead, last two two-week-ahead, last three-week-ahead). This is done separately for each time series to be predicted. Missing scores are imputed by the worst score achieved by any model for the respective target, meaning that irregularly submitted models will be penalized and receive less weight.

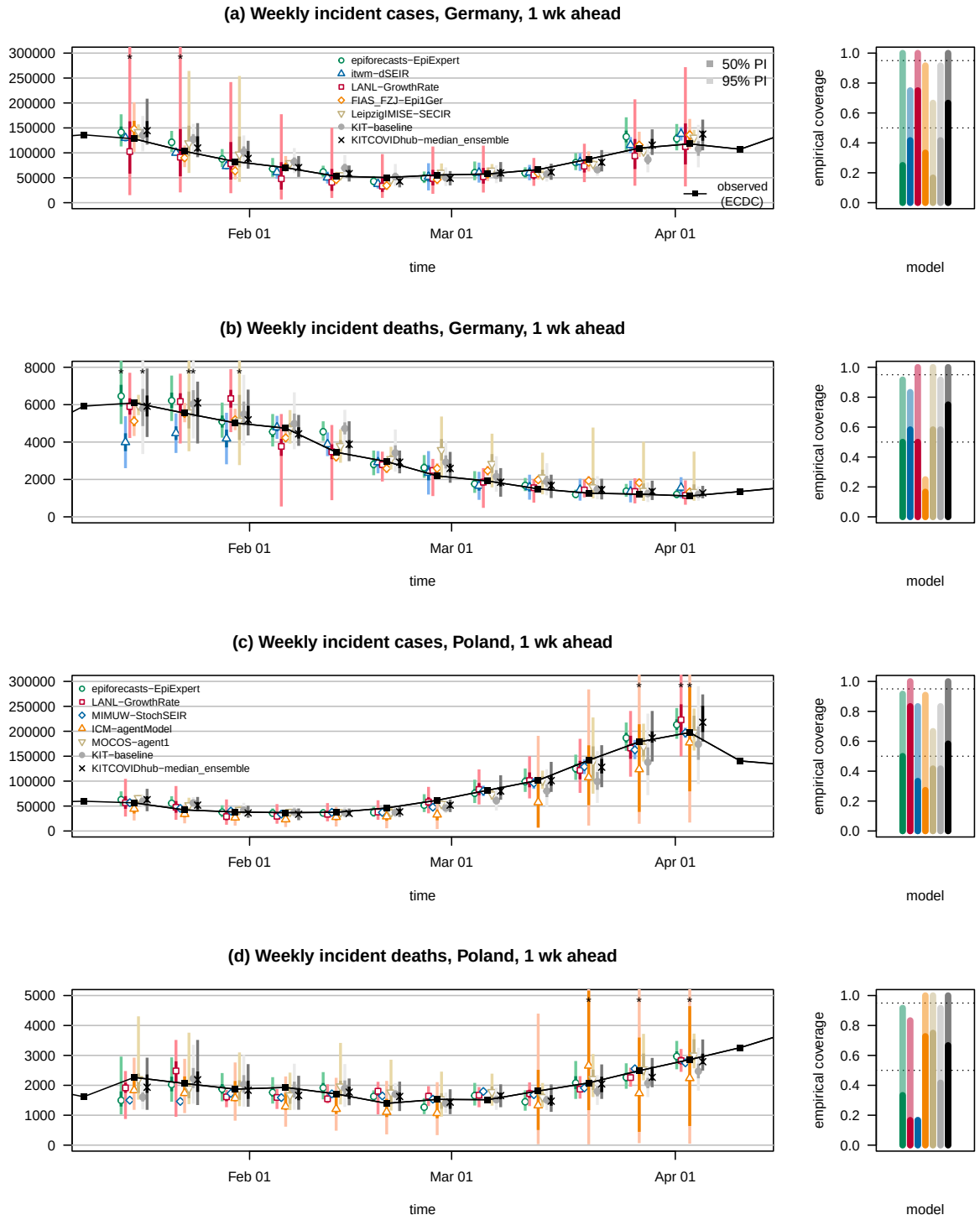
In the study protocol, the median ensemble was defined as our primary ensemble approach[11] as it can be assumed to be more robust to occasional misguided forecasts (e.g., due to technical errors). We therefore display this version in all figures and focus our discussion on it. Note that all forecast aggregations are performed directly at the level of quantiles rather than density functions as in other work.[26] This approach is referred to as Vincentization (in reference to Vincent,[27] see e.g., Busetti[28]). A broader discussion of Vincentization approaches and their application to epidemiological forecasts, including numerous other weighting schemes, can be found in recent works by Taylor and Taylor[29] and Ray et al.[30] Notably, Taylor and Taylor[29] used a similar inverse score weighting approach and found it to perform well in a re-analysis of forecasts from the US COVID-19 Forecast Hub. In this context we note that our inverse-WIS ensemble does not involve any estimation or optimization of weights, but simply uses the inverse of an average of past scores as heuristic weights. A more flexible approach with one tuning parameter estimated from the data has been used in Ray et al.[30]

There were no formal inclusion criteria other than completeness of the submitted set of 23 quantiles. The Forecast Hub team did, however, occasionally exclude forecasts with highly implausible central tendency or degree of dispersion manually. These exclusions have been documented in the Forecast Hub platform.

Results

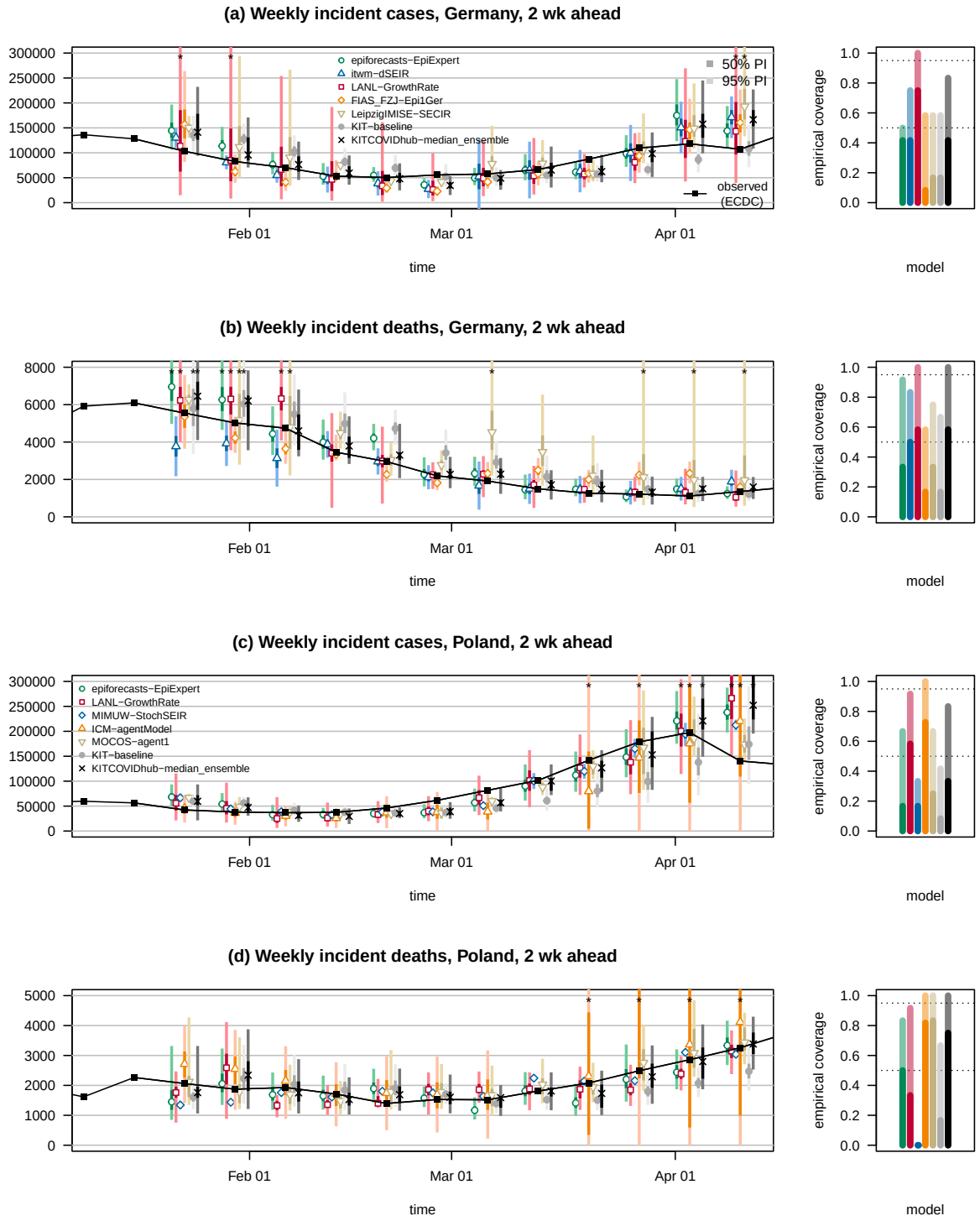
Figures 2 and 3 show the forecasts made by the median ensemble (KIT-median_ensemble; our pre-specified main ensemble approach; see Materials and Methods); a naïve model always using the last observed value as the expectation for the following weeks (KIT-baseline); and five contributed models with above-average overall performance across locations and targets (i.e., quantities to be predicted). The forecasts are probabilistic, and we display the 50% and 95% prediction intervals (PIs) along with the respective median. Forecasts by the remaining teams are illustrated in Supplementary Figures S1 and S2, and forecasts at horizons of three and four weeks are shown in Supplementary Figures S3–S6. In the following, we discuss the performance of these forecasts, starting with a formal statistical evaluation before directing attention to the behaviour at inflection points.

Figure 2: One-week ahead forecasts of cases and deaths from COVID-19 in Germany and Poland.



One-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. The figure shows forecasts from a baseline model, the median ensemble of all submissions and a subset of submitted models with above-average performance. The black line shows observed data. Colored points represent predictive medians, dark and light bars show 50% and 95% prediction intervals, respectively. Asterisks mark intervals exceeding the upper plot limit. The remaining submitted models are displayed in Supplementary Figure S1. The right column shows the empirical coverage rates of the different models. The dark and light bars represent the proportion of cases where the 50% and 95% prediction intervals, respectively, contained the observed values. Dotted lines show the desired nominal levels 0.5 and 0.95.

Figure 3: Two-week ahead forecasts of cases and deaths from COVID-19 in Germany and Poland.



Two-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. The figure shows forecasts from a baseline model, the median ensemble of all submissions and a subset of submitted models. The remaining submitted models are displayed in Supplementary Figure S2. The black line shows observed data. Colored points represent predictive medians, dark and light bars show 50% and 95% prediction intervals, respectively. The right column shows the empirical coverage rates of the different models. See caption of Figure 2 for a detailed explanation of plot elements.

Formal evaluation, January–April 2021. Table 2 and Figure 4 summarize the performance of the submitted, baseline and ensemble models over the twelve-week study period. Performance is measured via the average weighted interval score (WIS, see Methods section) and the mean absolute error of the predictive median. For both measures lower values indicate better predictive performance. We here show the average scores on the absolute scale, where they can be interpreted as the average distance between the observed and predicted value (the WIS taking into account forecast uncertainty). A summary table of relative scores standardized by the performance of the naïve KIT-baseline model is available in Supplementary Table S1. The WIS can moreover be decomposed into components representing underprediction, forecast spread and overprediction (see Methods), which we show in Supplementary Figure S7. Detailed results in tabular form at horizons of three and four weeks ahead can be found in Supplementary Table S2. As specified in the study protocol, we also provide results for cumulative cases and deaths (Supplementary Tables S3 and S4) and based on JHU rather than RKI/MZ data (Supplementary Tables S5 and S6; evaluation against JHU data leads to slightly higher WIS and absolute errors, but quite similar relative performance of models). A graphical display of individual scores can be found in Supplementary Figure S8.

Both for incident cases and deaths, a majority, but not all models outperformed the naïve baseline model KIT-baseline (a model outperforms the baseline for a given target whenever its bar in Figure 4 does not reach into the grey area). As one would expect, the performance of all models considerably deteriorated for longer forecast horizons. The pre-specified median ensemble was consistently among the best-performing methods, outperforming most individual forecasts for all targets. The KITCOVIDhub-inverse_wis_ensemble, which is an attempt to weigh member models based on recent performance, does not yield any clear benefits over the unweighted median and mean ensembles. As can be seen from Supplementary Figures S9 and S10, the weights fluctuate substantially, implying that the relative performance of different models may be too variable for performance-based weights to pay off. The KIT-extrapolation_baseline model shows quite reasonable relative performance for cases in both countries. Given the relatively long stretches of continued upward or downward trends in cases, this simple heuristic was not easy to beat and is rather close to the performance of the ensemble forecasts. For deaths, too, there are rather clear trends over the study period. Nonetheless, the different ensemble forecasts achieve substantial improvements over KIT-extrapolation_baseline, meaning that the deviations from the previous trends were predicted with some success.

The most striking cases of individual models outperforming the ensemble occurred for longer-range case forecasts in Poland. Here, the two microsimulation models MOCOS-agent1 and ICM-agentModel performed considerably better. These two models were arguably among the ones which were most meticulously tuned to the specific national context. It seems that this yielded benefits for longer horizons, while at shorter horizons the ensemble and some considerably simpler models were at least on par (the best performance at the one week horizon being achieved by the compartmental model MIMUW-StochSEIR).

There were considerable differences in the forecast uncertainty of the different models. This can be seen from the quite variable forecast interval widths in Figures 2 and 3, and resulted in large differences in the empirical coverage rates of 50% and 95% prediction intervals (Table 2 and right column in the aforementioned figures). The ensemble methods performed quite favourably in terms of coverage, typically with slight undercoverage (i.e., prediction intervals cover the observations less frequently than intended) for cases and slight overcoverage (intervals cover more often than intended) for deaths. The differences in forecast dispersion are also reflected by the components of the weighted interval score shown in Supplementary Figure S7 (see Materials and Methods for an explanation of the decomposition). Some models, most strikingly ITWW-county_repro, issued very sharp predictions, leading to very small dispersion components of the weighted interval score (the darkest block in the middle of the stacked bar). In turn, this model received rather large penalties for both over- and underprediction. Other models, like LANL-GrowthRate, epiforecasts-EpiNow2 and ICM-agentModel issued comparatively wide forecasts, leading to WIS values with large dispersion components. While there is no clear rule on what the score decomposition of an ideal forecast should look like, comparisons of the components provide useful indications on how to improve a model (e.g., the ITWW-county_repro model might benefit from widening the uncertainty intervals).

A subset of models also provided forecasts at the subnational level (states in Germany, voivodeships in Poland). Table 3 provides a summary of the respective results at the one and two week horizons (results for three and four weeks can be found in Supplementary Table S7). Despite the rather low number of available models, the ensembles generally achieved improvements over the individual models and, with exceptions for case forecasts in Germany, clearly outperformed the baseline model KIT-baseline. The mean WIS values are lower for the regional forecasts than for the national-level forecasts in Table 2 primarily because the

numbers to be predicted are lower at the regional level; the WIS – like the absolute error – scales with the order of magnitude of the predicted quantity and cannot be compared directly across different forecasting tasks. Coverage of the ensemble forecasts was close to the nominal level for deaths and somewhat lower for cases. Note that in this comparison part of the forecasts from the FIAS_FZJ-epi1Ger model were created retrospectively (using only the data available up to the forecast date) as the team only started issuing forecasts for all German federal states on 22 February 2021.

As specified in the study protocol,[11] we also report evaluation results at the national level pooled across the two study periods for those models which covered both. These are summarized in Supplementary Tables S8 and S9. For deaths, ensemble forecasts clearly outperformed individual models, the four-week-ahead horizon in Poland being the only one at which an individual model (epiforecasts-EpiExpert) meaningfully outperformed the pre-specified median ensemble. While most contributed and baseline models were somewhat overconfident, the ensemble showed close to nominal coverage even at the four-week-ahead horizon. For cases, the median ensemble achieved good relative performance (comparable to the best individual models) one and two weeks ahead, but was outperformed by a number of other models at three and four weeks. Notably, it failed to beat the naïve last-observation-carried-forward model KIT-baseline. Its coverage of prediction intervals was acceptable one week ahead, but substantially below nominal at higher horizons (e.g., 13/19 and 10/19 four weeks ahead in Germany and Poland, respectively, at the 0.95 level), which reflects the severe difficulties in predicting cases in Fall 2020 as discussed in Bracher et al.[16]

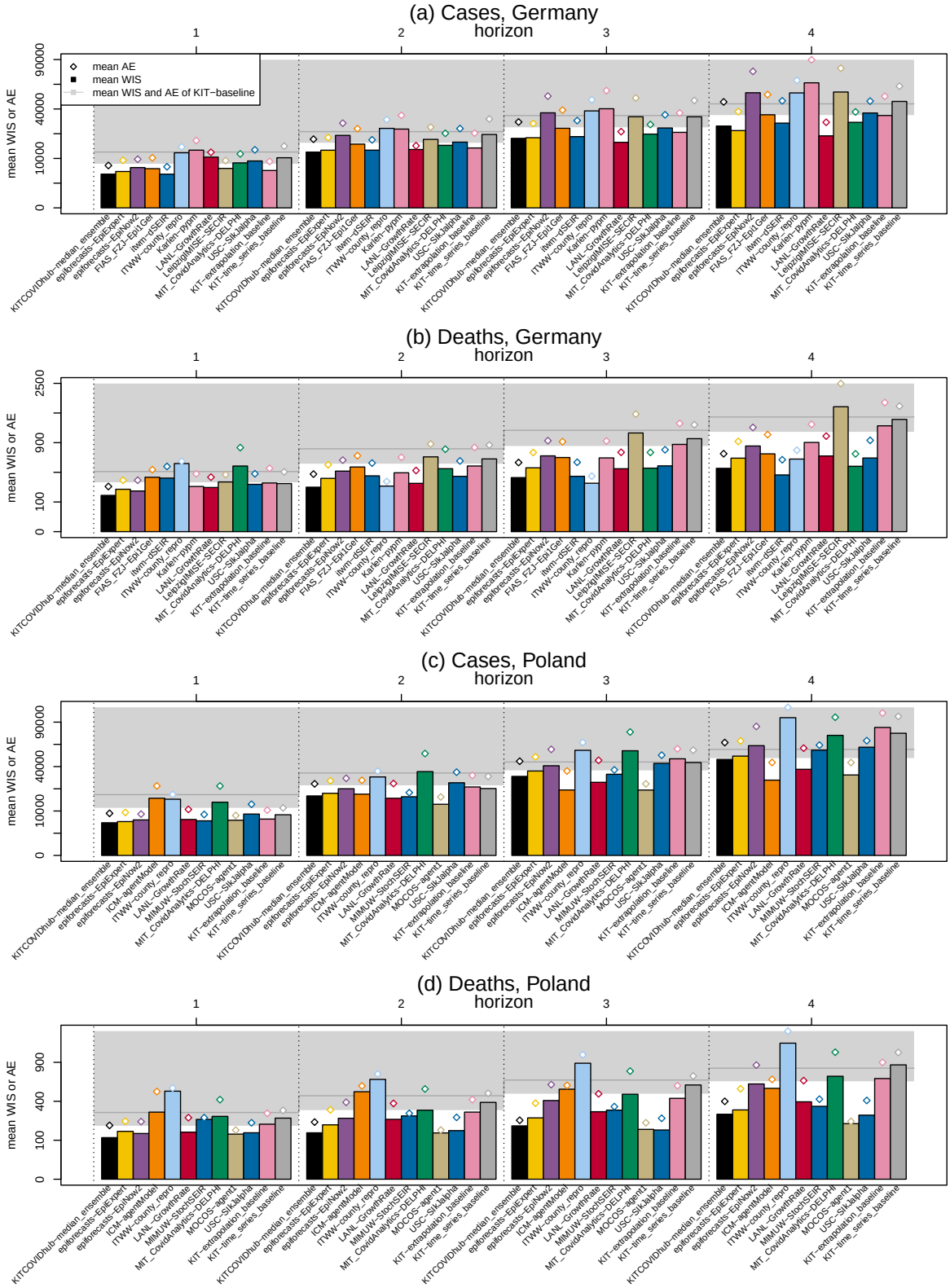
Behaviour at inflection points. From a public health perspective, there is often a specific interest in how well models anticipated major inflection points (changes in trend). We therefore discuss these instances separately. However, we note that, as will be detailed in the discussion, post-hoc conditioning of evaluation results on the occurrence of unusual events comes with important conceptual challenges.

Shift from wild type to B.1.1.7 variant. The renewed increase in cases in both Germany and Poland (third wave) in late February 2021 was due to the shift from the wild-type variant of the virus to the B.1.1.7 (or Alpha) variant, see Figure 1, panel (c) for estimated shares of the new variant over time. Given earlier observations about the spread of the B.1.1.7 variant in the UK[31] and Denmark, there was public discussion about the likelihood of a re-surgence, but there was considerable uncertainty about the timing and strength (see e.g., a German newspaper article[32] from early February 2021). This was largely due to the limited availability of representative sequencing data. In certain regions of Germany, specifically the city of Cologne[33] and the state of Baden-Württemberg,[34] large-scale sequencing had been adopted by late January, but results were considered difficult to extrapolate to the whole of Germany. An updated RKI report[35] on virus variants from 10 February 2020 described a "continuous increase in the share of the VOC B.1.1.7", but cautioned that the data were "subject to biases, e.g., with respect to the selection of samples to sequence" (our translation).

Given the limited available data, and the fact that many approaches had not been designed to accommodate multiple variants, only two of the teams submitting forecasts for Germany opted to account for this aspect (a question which was repeatedly discussed during coordination calls). These exceptions were the Karlen-pypm and LeipzigIMISE-SECIR models, which starting from 1 March 2021 explicitly accounted for the presence of two variants. As a result, most models did not anticipate the change in trend well and only reacted implicitly once the change became apparent in the data on 27 February 2021. Figure 5 shows the case forecasts of all submitted models and the median ensemble from 15 February, 22 February and 1 March 2021. We also show the two short time series of shares of the B.1.1.7 variant available from Robert Koch Institute at the respective prediction time points.

The ITWW-county_repro model was the only one to anticipate a change in trend on 15 February (though slower than the observed one), and adapted quickly to the upward trend in the following week. This model extrapolates recently observed growth or decline at the county-level and aggregates these fine-grained forecasts to the state or national level. Therefore it may have been able to catch a signal of renewed growth, as a handful of German states had already experienced a slight increase in cases in the previous week (e.g., Thuringia and Saxony-Anhalt, see panel (b) of Supplementary Figure S11). However, as illustrated in panel (a) of the same Figure, the ITWW model had also predicted turning points earlier during the same phase of decline in cases, and might generally have a tendency to produce such patterns. Another noteworthy observation in this context is the change in the predictions of the Karlen-pypm model. After the extension of the model to account for the B.1.1.7 variant on 1 March, its forecasts changed from the most optimistic to the most pessimistic among all included models (panels b and c of Figure 5). The other model including

Figure 4: Formal evaluation results in terms of mean weighted interval scores.



Average weighted interval scores (bars) and absolute errors (diamonds) achieved by models across countries, targets and forecast horizons (12 weekly forecasts). The bottom end of the grey area represents the mean WIS of the baseline model KIT-baseline and the grey horizontal line its mean absolute error. Values are shown on a square-root scale to enhance readability. Only models covering all four horizons are shown.

Table 2: Forecast evaluation for Germany and Poland (incidence scale, based on RKI/MZ data).

	Germany															
	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
Model	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	9,252	5,415	0.25	1.00	20,233	13,607	0.42	0.50	300	204	0.50	0.92	509	323	0.33	0.92
epiforecasts-EpiNow2	9,676	6,644	0.67	0.83	29,348	21,478	0.58	0.67	300	188	0.75	0.92	581	417	0.67	0.75
FIAS_FZJ-Epi1Ger	10,218	6,294	0.33	0.92	25,662	16,621	0.08	0.58	436	336	0.17	0.25	655	475	0.17	0.58
IHME-CurveFit									516				656			
Imperial-ensemble2									*193	*136	0.80	0.90				
itwm-dSEIR	6,905	4,644	0.42	0.75	18,935	13,626	0.42	0.75	483	326	0.58	0.83	534	354	0.50	0.83
ITWW-county_repro	15,223	12,418	0.08	0.25	31,836	25,851	0.00	0.17	564	527	0.00	0.00	286	236	0.08	0.25
Karlen-pypm	18,532	13,629	0.50	0.92	35,010	25,385	0.25	0.83	380	232	0.42	0.92	628	394	0.08	0.83
LANL-GrowthRate	12,623	10,542	0.75	1.00	15,797	13,945	0.75	1.00	338	222	0.50	1.00	425	265	0.58	1.00
LeipzigIMISE-SECIR	9,161	6,376	0.17	0.67	26,650	19,185	0.17	0.58	370	281	0.58	1.00	874	636	0.33	0.75
MIT_CovidAnalytics-DELPHI	*11,910	* 8,277	0.55	0.91	*22,734	*16,006	0.36	0.73	803	490	0.42	0.92	773	451	0.58	1.00
SDSC_ISG-TrendModel	7,861								436							
USC-SIkJalpha	13,766	9,001	0.25	0.83	25,730	17,681	0.17	0.58	381	255	0.50	0.83	568	348	0.25	0.83
KIT-baseline	12,756	7,953	0.42	0.92	23,785	17,330	0.17	0.58	411	277	0.58	0.92	780	525	0.17	0.67
KIT-extrapolation_baseline	8,823	5,715	0.50	1.00	22,858	14,679	0.33	0.75	456	269	0.33	1.00	806	490	0.33	0.83
KIT-time_series_baseline	15,583	10,281	0.25	0.75	32,306	22,026	0.25	0.67	406	263	0.67	1.00	851	601	0.50	0.92
KITCOVIDhub-inverse_wis_ensemble	8,586	5,294	0.58	1.00	22,000	13,824	0.50	0.83	216	149	0.75	1.00	307	207	0.75	1.00
KITCOVIDhub-mean_ensemble	8,377	5,277	0.75	1.00	21,825	13,662	0.50	0.92	220	152	0.58	1.00	346	219	0.75	1.00
KITCOVIDhub-median_ensemble	7,344	4,660	0.67	1.00	19,296	12,734	0.42	0.83	232	150	0.75	1.00	376	225	0.58	1.00

	Poland															
	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
Model	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	7,500	4,553	0.50	0.92	25,316	17,408	0.17	0.67	208	137	0.33	0.92	287	181	0.50	0.83
epiforecasts-EpiNow2	7,928	5,906	0.58	0.92	29,762	22,098	0.42	0.83	184	119	0.58	1.00	340	228	0.58	1.00
ICM-agentModel	*23,011	*15,824	0.27	0.91	*26,694	*18,098	0.73	1.00	*488	*294	0.73	1.00	*605	*507	0.82	1.00
IHME-CurveFit									374				520			
Imperial-ensemble2									*188	*138	0.30	0.70				
ITWW-county_repro	20,054	17,364	0.17	0.25	36,651	31,445	0.17	0.50	589	551	0.00	0.00	784	711	0.00	0.00
LANL-GrowthRate	8,129	5,787	0.83	1.00	23,269	15,240	0.58	0.92	229	137	0.17	0.83	347	216	0.33	0.92
MIMUW-StochSEIR	5,705	4,028	0.33	0.83	17,642	15,347	0.17	0.33	237	224	0.17	0.17	288	267	0.00	0.00
MIT_CovidAnalytics-DELPHI	*22,344	*12,912	0.20	0.90	*49,687	*33,033	0.10	0.70	*393	*244	0.55	1.00	*520	*296	0.36	1.00
MOCOS-agent1	5,173	4,978	0.42	0.67	15,022	11,380	0.25	0.67	158	132	0.75	1.00	203	149	0.83	1.00
SDSC_ISG-TrendModel	6,323								265							
USC-SIkJalpha	10,404	6,919	0.33	0.83	32,822	24,436	0.17	0.50	206	133	0.33	0.92	266	168	0.42	0.92
KIT-baseline	16,407	9,736	0.42	0.83	32,182	22,709	0.08	0.42	258	167	0.42	0.92	416	275	0.17	0.67
KIT-extrapolation_baseline	9,448	5,992	0.50	0.92	29,638	22,165	0.25	0.58	269	190	0.58	0.83	404	284	0.42	0.83
KIT-time_series_baseline	10,784	7,787	0.75	0.83	30,359	21,510	0.50	0.75	300	232	0.67	0.67	467	362	0.50	0.58
KITCOVIDhub-inverse_wis_ensemble	7,319	4,689	0.50	1.00	23,418	15,580	0.42	0.83	150	111	0.75	1.00	197	144	0.75	1.00
KITCOVIDhub-mean_ensemble	6,866	4,784	0.75	1.00	23,673	15,573	0.33	0.83	141	114	0.75	1.00	173	152	0.92	1.00
KITCOVIDhub-median_ensemble	7,130	4,403	0.58	1.00	23,027	16,241	0.33	0.83	162	103	0.67	1.00	193	137	0.75	1.00

Summaries are based on 12 weekly forecasts per target.

Abbreviations: $C_{0.5}$, $C_{0.95}$: coverage rates of the 50% and 95% prediction intervals; AE: mean absolute error; WIS: mean weighted interval score.

*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: Forecast evaluation at the regional level, Germany and Poland (incidence scale, RKI/MZ data).

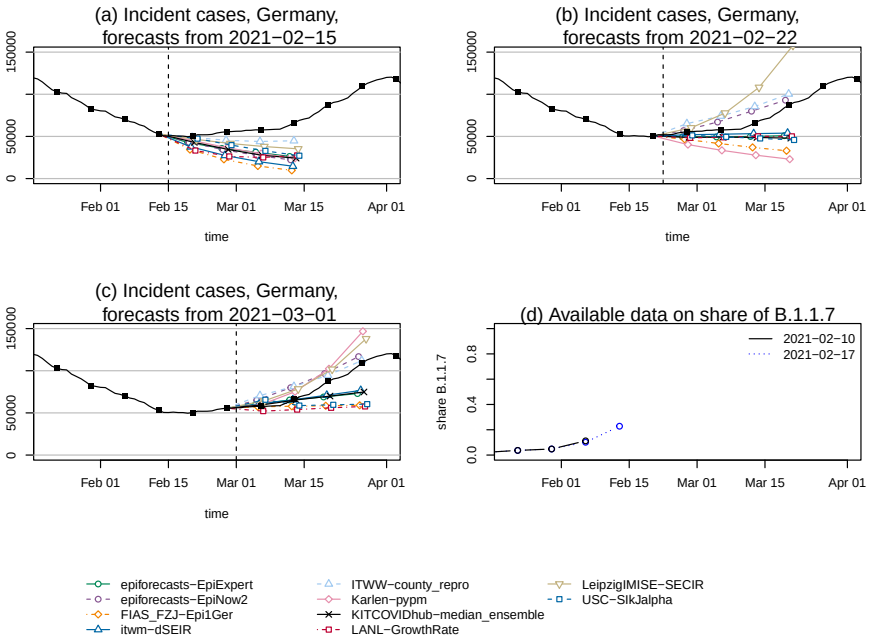
	Germany															
	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
Model	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiNow2	694	448	0.52	0.90	1,663	1,120	0.39	0.77	36	24	0.54	0.92	65	43	0.39	0.84
FIAS_FZJ-Epi1Ger	801	518	0.30	0.76	1,732	1,173	0.25	0.66	41	34	0.11	0.30	54	41	0.12	0.45
IHME-CurveFit									47				53			
ITWW-county_repro	996	671	0.36	0.73	2,053	1,392	0.40	0.78	41	34	0.15	0.46	28	20	0.26	0.69
Karlen-pypm	1,251	804	0.65	0.95	2,300	1,432	0.48	0.92	35	22	0.62	0.94	48	28	0.47	0.95
LeipzigIMISE-SECIR			0.08	0.58			0.17	0.33			0.08	0.42			0.08	0.42
USC-SIkJalpha	941	631	0.45	0.86	1,614	1,094	0.32	0.74	32	21	0.46	0.91	43	28	0.33	0.83
KIT-baseline	859	544	0.42	0.89	1,536	1,085	0.28	0.65	38	25	0.51	0.92	57	38	0.35	0.78
KIT-extrapolation_baseline	785	504	0.47	0.89	1,616	1,073	0.34	0.76	49	31	0.46	0.88	76	50	0.43	0.85
KIT-time_series_baseline	1,010	665	0.40	0.80	1,954	1,320	0.28	0.68	59	36	0.36	0.92	93	59	0.33	0.84
KITCOVIDhub-inverse_wis_ensemble	868	523	0.49	0.97	1,744	1,098	0.38	0.89	28	18	0.56	0.97	39	25	0.50	0.94
KITCOVIDhub-mean_ensemble	848	515	0.50	0.97	1,693	1,044	0.43	0.91	28	18	0.56	0.97	41	26	0.51	0.96
KITCOVIDhub-median_ensemble	769	485	0.57	0.95	1,661	1,028	0.42	0.90	29	19	0.57	0.94	36	23	0.49	0.95

	Poland															
	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
Model	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-EpiNow2	632	426	0.58	0.89	2,081	1,500	0.41	0.78	27	18	0.49	0.86	47	33	0.44	0.82
ITWW-county_repro	1,351	956	0.36	0.70	2,532	1,867	0.40	0.74	45	38	0.12	0.31	56	44	0.19	0.42
USC-SIkJalpha	853	622	0.40	0.90	1,841	1,305	0.25	0.64	20	13	0.51	0.95	27	18	0.45	0.90
KIT-baseline	1,085	647	0.34	0.84	2,028	1,451	0.15	0.52	26	17	0.51	0.91	34	23	0.36	0.79
KIT-extrapolation_baseline	702	457	0.47	0.89	2,069	1,485	0.34	0.70	29	19	0.46	0.89	39	25	0.43	0.89
KIT-time_series_baseline	804	562	0.53	0.83	2,123	1,459	0.42	0.73	32	21	0.41	0.80	44	31	0.32	0.70
KITCOVIDhub-inverse_wis_ensemble	658	399	0.60	0.98	1,786	1,155	0.43	0.82	19	12	0.53	0.93	25	16	0.51	0.93
KITCOVIDhub-mean_ensemble	604	380	0.61	0.99	1,663	1,061	0.43	0.88	19	13	0.53	0.93	26	17	0.52	0.94
KITCOVIDhub-median_ensemble	615	385	0.64	0.97	1,801	1,125	0.42	0.83	20	13	0.56	0.93	26	17	0.50	0.94

Results are averaged over the different regions (16 states in Germany, 16 voivodeships in Poland).

Abbreviations: $C_{0.5}$, $C_{0.95}$: coverage rates of the 50% and 95% prediction intervals; AE: mean absolute error; WIS: mean weighted interval score.
*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.

Figure 5: Case forecasts in Germany preceding the upward trend change in March 2022.



Point forecasts of cases in Germany, as issued on (a) 15 February, (b) 22 February and (c) 1 March 2021. These dates, shown as vertical dashed lines, mark the start of a renewed increase in overall case counts due to the new variant of concern B.1.1.7. Panel (d): Data by RKI on the share of the B.1.1.7 variant as available on the different forecast dates (the next data release by RKI occurred on 3 March). The models Karlen-pypm and LeipzigIMISE-SECIR accounted for the presence of multiple variants from 1 March onwards.

variant data, LeipzigIMISE-SECIR, likewise was among the first to adopt an upward trend.

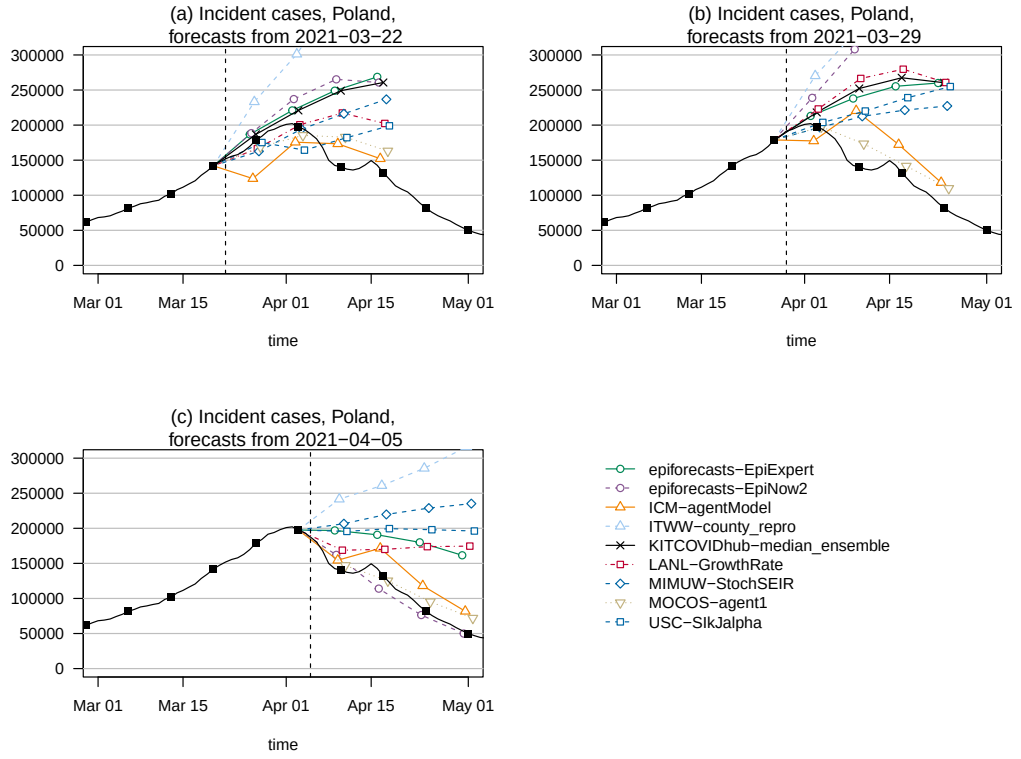
In Poland, availability of sequencing data was very limited during our study period; the GISAID database[19] only contained 2271 sequenced samples for Poland by 29 March 2021.[18] Nonetheless, the ICM-agentModel and MOCOS-agent1 models explicitly took the presence of a new variant into account to the degree possible. Again, the ITWW-county_repro model was the first to predict a change in overall trends (in this case without having predicted turning points already in the preceding weeks; see Supplementary Figure S1).

Peak of the third wave (cases). In Poland, the third wave reached its peak in the week ending on 3 April 2021. Despite the fact that it coincided with the Easter weekend and thus somewhat unclear data quality, this turnaround was predicted quite well by two Poland-based teams, MOCOS-agent1 and ICM-agentModel. As can be seen from Figure 6, the trajectory of these two models differed substantially from those of most other models, including the ensemble, which predicted a sustained increase. This successful prediction of the turning point was in large part responsible for the good relative performance of MOCOS-agent1 and ICM-agentModel at longer horizons (Table 2). In retrospective discussions, the respective teams noted that the tightening of non-pharmaceutical interventions (NPIs) on 27 March (which they had anticipated) in combination with possible seasonal effects had led them to expect a downward turn.

For Germany, the peak of the third wave occurred only after the end of our pre-specified study period, but we note that numerous models showed strong overshoot as they expected the upward trend to continue. The exact mechanisms underlying the turnaround remain poorly understood. A new set of restrictions referred to as the Bundesnotbremse in German (federal emergency break) was introduced too late to explain the change on its own.

Changes in trend of deaths. In Germany, the study period coincided almost perfectly with a prolonged period of decline in deaths. In Figure 7, panels (a) and (b) show the behaviour of the median ensemble at

Figure 6: Case forecasts in Poland surrounding the peak in April 2022.



Point forecasts of cases in Poland from (a) 22 March, (b) 29 March and (c) 5 April 2021, surrounding the peak week. In each panel, the date at which forecasts were created is marked by a dashed vertical line. The models ICM-agentModel and MOCOC-agent1 anticipated the trend change correctly, while the remaining models show more or less pronounced overshoot.

the beginning and end of this phase. The ensemble had already anticipated a downward turn on 4 January, two weeks before it actually occurred. Following the unexpected strong increase in the following week, it went to extending the upward tendency, before switching back to predicting a turnaround. It seems likely that the irregular pattern in late December and early January is partly due to holiday effects in reporting, and forecast models may have been disturbed by this aspect.

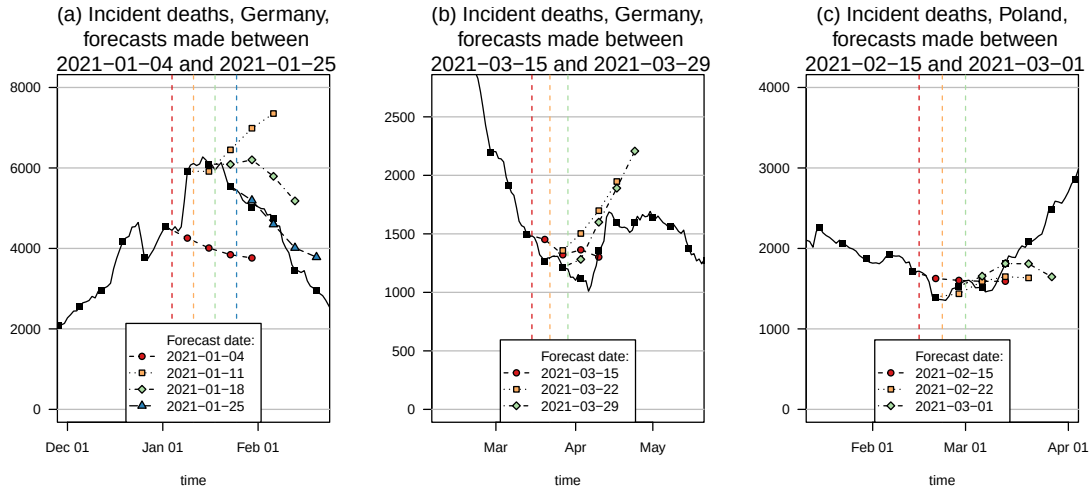
At the end of the downward trend in late March, the ensemble again anticipated the turnaround to arrive earlier than it did, and predicted a more prolonged rise than was observed. Nonetheless, in both cases the ensemble to some degree anticipated qualitative change, and the observed trajectories were well inside the respective 95% prediction intervals (with the exception of the forecast from 4 January; however, this forecast had prospectively been excluded from the analysis as we anticipated reporting irregularities).

In Poland, deaths started to increase in early March after a prolonged period of decay. As can be seen in panel c of Figure 7, the median ensemble had anticipated this change (22 February 2021), but in terms of its point forecast did not initially expect a prolonged upward trend as later observed. Nonetheless, the observed trajectory was contained in the relatively wide 95% prediction intervals (Figures 2 and 3).

Discussion

We presented results from the second and final part of a pre-registered forecast evaluation study conducted in Germany and Poland (January–April 2021). During the period covered in this paper, ensemble approaches yielded very good performance relative to contributed individual models and baseline models. The majority of contributed models was able to outperform a simple last-observation-carried-forward model for most

Figure 7: Death forecasts preceding trend changes.



Point forecasts of the median ensemble during changing trends in deaths. Panel (a): Downward turn in Germany, January 2021. Panel (b): Upward turn in Germany, March 2021. Panel (c): Upward turn in Poland, February/March 2021. Different colours and point/line shapes represent forecasts made at distinct time points (marked by dashed vertical lines).

targets and forecast horizons up to four weeks.

The results in this manuscript differ in important aspects from those for our first evaluation period (October–December 2020), when most models struggled to meaningfully outperform the KIT-baseline model for cases. Fall 2020 was characterized by rapidly changing non-pharmaceutical intervention measures, making it hard for models to anticipate the case trajectory. Pooled across both study periods, we found ensemble forecasts of deaths to yield satisfactory reliability and clear improvements over baseline models. For cases, however, coverage was clearly below nominal from the two-week horizon onward, and in terms of mean weighted interval scores the ensemble failed to outperform the KIT-baseline model three and four weeks ahead. This strengthens our previous conclusion[16] that meaningful case forecasts are only feasible at very short horizons. It also agrees with recent results from the US COVID-19 Forecast Hub,[36] which led the organizers to temporarily suspend ensemble case forecasts beyond the one-week horizon.

The differences between our two study periods illustrate that performance relative to simple baseline models is strongly dependent on how good a fit these are for a given period. Cases in Germany plateaued during November and early December 2020, making the last-observation-carried-forward strategy of KIT-baseline difficult to beat. The second evaluation period was characterized by longer stretches of continued upward or downward trends, making it much easier to beat that baseline. In this situation, however, many models did not achieve strong improvements over the extrapolation approach KIT-extrapolation_baseline. Ideally one would wish complex forecast models to outperform each of these different baseline models. However, there are many ways of specifying a simple baseline,[37] and post-hoc at least one of them will likely be in acceptable agreement with the observed trajectory. While the choice of the most meaningful reference remains subject to debate, we believe that the use of a small set of pre-specified baselines as in the present study is a reasonable approach.

An observation made for both the first and the second part of our study is that predicting changing trends in cases is very challenging; turnarounds in death counts are less difficult to anticipate. This finding is shared by works on real-time forecasts of COVID-19 in the UK[38] and the US.[39] To interpret these insights we note that, in principle, there are two ways of forecasting epidemiological time series:

(i) Applying a mechanistic model to project future spread based on recent trends and other relevant factors like NPIs, population behaviour, spread of different variants or vaccination. Models can then predict trend changes based on classical epidemiological mechanisms (depletion of susceptibles) or observed/anticipated changes in surrounding factors, which depending on the model may be treated as exogenous or endogenous.

(ii) Establishing a statistical relationship (often with a mechanistic motivation) to a leading indicator, i.e. a data stream which is informative on the trajectory of the quantity of interest, but available earlier. Changes in the trend of the leading indicator can then help anticipate future turning points in the time series of interest.

Death forecasts belong into the realm of category (ii), with cases and hospitalizations serving as leading indicators. This prediction task has been addressed with considerable success. Case forecasts, on the other hand, typically are based on approach (i), which largely reduces to trend extrapolation, unless models are carefully tuned to changing NPIs (see Table 1). Theoretical arguments on the limited predictability of turning points in such curves have been brought forward,[40,41] and empirical work including ours confirms that this is a very difficult task. While the success of the two microsimulation models MOCOS-agent1 and ICM-agentModel in anticipating the downward turn in cases in Poland remains a rather rare exception, it shows that careful mechanistic modelling combined with in-depth knowledge of national specificities has the potential to anticipate the impact of changing NPIs. As both groups heavily drew from experience from past NPIs in Poland, there is hope that predictions of the effects of NPIs will further improve as experience accumulates. An alternative strategy to improve case forecasts would be to identify appropriate leading indicators. These could for instance be trajectories in other countries[42] or additional data streams on e.g., mobility, insurance claims or web searches. However, the benefits of such data for short-term forecasting thus far have been found to be modest.[43] Changes in dominant variants may make changes in overall trends predictable as they arise from the superposition of adverse but stable trends for the different variants. The availability of sequencing data has improved considerably since our study period, and we consider the extension of models to accommodate multiple strains a key step towards improved prediction of trend changes. Other relevant aspects include seasonal effects, which during our study period remained poorly understood due to limited historical data, and population immunity. As more data on seroprevalence become available, predictability of saturation effects may increase, though this will likely be complicated by the further evolution of the pathogen.

Another difficulty of case forecasts is incomplete case ascertainment, which must be assumed to vary over time.[9,44] As a consequence, data can be difficult to compare across different phases of the pandemic, and modellers often choose to only use a recent subset of the available data to calibrate their models. While data on testing volumes and test positivity rates are available, estimation of the reporting fractions and anticipation of its future development is challenging. Even if models correctly reflect underlying epidemic dynamics, this may thus not translate to accurate forecasts of the observed number of confirmed cases. This is a limitation of the considered forecasts and their evaluation, which inherit the difficulties of the underlying truth data sources. Nonetheless, we argue that a distinguishing feature of forecasts is that they refer to observable quantities, and forecasters should take into account all relevant aspects of the system producing them. Indeed, many of the considered models (e.g., MOCOS-agent1 and FIAS_FZJ-Epi1Ger) attempt to reconstruct the underlying infection dynamics, which are then linked to the number of reported cases via time-varying reporting probabilities.

We have extensively discussed the difficulties models encountered at turning points. In the aftermath of such events, epidemic forecasts typically receive increased attention in the general media (e.g., following the rapid downward turn in cases in Germany in May 2021[45]). While important from a subject-matter perspective, this is not without problems from a formal forecast evaluation standpoint. Major turning points are rare events and as such difficult to forecast. Focusing evaluation on solely these instances will benefit models with a strong tendency to predict change, and adapting scoring rules to emphasize these events in a principled way is not straightforward. This problem is known as the forecaster's dilemma[46] in the literature and likewise occurs in, e.g., economics and meteorology (see illustrations in Table 1 from Lerch et al[46]). An interesting question for future work is whether turning points are preceded by stronger disagreement between models, which might then serve as an alert; or whether, on the contrary, trend changes are followed by increased disagreement. Especially the latter question has received considerable attention in economic forecasting.[47]

In this paper we only applied unweighted ensembles and a heuristic, rather unflexible weighting scheme based directly on past average performance. More sophisticated weighting schemes have been explored in[29] and[39] using data from the US COVID-19 Forecast Hub. Their results indicate that when some contributing forecasters have a stable record of good performance, giving these more weights can result in improved performance. In particular, restricting the ensemble to a set of well-performing models may be beneficial,

a strategy employed in the so-called relative WIS weighted median ensemble[39] used by the US COVID-19 Forecast Hub since November 2021.

The present paper marks the end of the German and Polish COVID-19 Forecast Hub as an independently run platform. In April 2021, the European Center for Disease Prevention and Control (ECDC) announced the launch of a European COVID-19 Forecast Hub,[4] which has since attracted submissions from more than 30 independent teams. The German and Polish COVID-19 Forecast Hub has been synchronized with this larger effort, meaning that all forecasts submitted to our platform are forwarded to the European repository, while forecasts submitted there are mirrored in our dashboard. In addition, we still collect regional-level forecasts, which are not currently covered in the European Forecast Hub. The adoption of the Forecast Hub concept by ECDC underscores the potential of collaborative forecasting systems with combined ensemble predictions as a key output, along with continuous monitoring of forecast performance. We anticipate that this closer link to public health policy making will enhance the usefulness of this system to decision makers. An important step will be the inclusion of hospitalization forecasts. Due to unclear data access, these had not been tackled in the framework of the German and Polish COVID-19 Forecast Hub, but have been added in the new European version.

Data availability

The forecast data generated in this study have been deposited in a GitHub repository (https://github.com/KITmetricslab/covid19-forecast-hub-de), with a stable Zenodo release available under accession code 5608390 https://zenodo.org/record/5608390#.YYFxdJso9H4. This repository also contains all truth data used for evaluation. Forecasts can be visualised interactively at https://kitmetricslab.github.io/forecasthub/. Should any further data be required to reproduce the results these can be obtained from the corresponding authors upon reasonable request.

Code availability

Codes to reproduce figures and tables are available at https://github.com/KITmetricslab/analyses_de_pl2, with a stable version at https://zenodo.org/record/5639514#.YYF1aZso9H4.[61] The results presented in this paper have been generated using the release preprint2 of the repository https://github.com/KITmetricslab/covid19-forecast-hub-de, see above for the link to the stable Zenodo release. The codes require the R packages colorspace (version 1.8-9),[62] plotrix (version 3.8-1),[63] xtable (version 1.8-4)[64] and zoo (version 1.8-9[65]).

Competing interests

The authors declare no competing interests.

References

[1] Ray, E. L. *et al.* Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. *medRxiv* (2020). URL https://www.medrxiv.org/content/early/2020/08/22/2020.08.19.20177493.

[2] Cramer, E. Y. *et al.* Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences* **119**, e2113561119 (2022).

[3] Borchering, R. K. *et al.* Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios – United States, April–September 2021. *Morbidity and Mortality Weekly Report* **70**, 719–724 (2021).

[4] Sherratt, K. *et al.* Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *medRxiv* (2022). URL https://www.medrxiv.org/content/early/2022/06/16/2022.06.16.22276024.

[5] McGowan, C. J. *et al.* Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports* **9**, 683 (2019).

[6] Del Valle, S. *et al.* Summary results of the 2014–2015 DARPA Chikungunya Challenge. *BMC Infectious Diseases* **18**, 245 (2018).

[7] Johansson, M. A. *et al.* An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences* **116**, 24268–24274 (2019).

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

[9] Arık, S. *et al.* A prospective evaluation of AI-augmented epidemiology to forecast COVID-19 in the USA and Japan. *npj Digital Medicine* **4**, 146 (2021).

[10] Dirnagl, U. Politikberatung, bis der Elefant mit dem Rüssel wackelt! *Laborjournal* **5/2021**, 22–24 (2021).

[11] Bracher, J., the German and Polish COVID-19 Forecast Hub Team & Participants. 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 (2020).

[12] Robert Koch Institut. CSV mit den aktuellen Covid-19 Infektionen pro Tag (Zeitreihe). Available online, https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74, last accessed on 18 August 2022. (2022).

[13] Polish Ministry of Health. Dane historyczne dla województw. Available online, https://www.arcgis.com/home/item.html?id=a8c562ead9c54e13a135b02e0d875ffb, last accessed on 18 August 2022. (2022).

[14] Johns Hopkins University Center for Systems Science and Engineering. COVID-19 Data Repository. Available online, https://github.com/CSSEGISandData/COVID-19, last accessed on 18 August 2022. (2022).

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

[16] Bracher, J. *et al.* A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nature Communications* **12**, 5173 (2021).

[17] Robert Koch Institut. Bericht zu Virusvarianten von SARS-CoV-2 in Deutschland, insbesondere zur Variant of Concern (VOC) B.1.1.7, 31 March 2021. Available at https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/Bericht_VOC_2021-03-31.pdf (2021).

[18] MI2 Data Lab, Warsaw University of Technology. Monitor of SARS-CoV-2 variants, version 2021-05-05 (2021). Available at https://monitor.crs19.pl/2021-05-05/poland/?lang=en.

[19] GISAID Initiative. Enabling rapid and open access to epidemic and pandemic virus data – tracking of variants (2021). Available at https://www.gisaid.org/hcov19-variants/.

[20] Hale, T. *et al.* A global panel database of pandemic policies (Oxford COVID-19 government response tracker). *Nature Human Behaviour* **5**, 529–538 (2021).

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

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

[23] Petropoulos, F. & Makridakis, S. Forecasting the novel coronavirus COVID-19. *PLOS ONE* **15**, e0231236 (2020).

[24] Hyndman, R. *et al.* forecast: Forecasting functions for time series and linear models. R package version 8.12.0, URL https://pkg.robjhyndman.com/forecast/. (2021).

[25] IHME COVID-19 Forecasting Team. Modeling COVID-19 scenarios for the United States. *Nature Medicine* **27**, 94–105 (2021).

[26] Reich, N. G. *et al.* A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences* **116**, 3146–3154 (2019).

[27] Vincent, S. The function of the viborissae in the behavior of the white rat. *Behavioral Monographs* **1**, 1–82 (1912).

[28] Busetti, F. Quantile aggregation of density forecasts. *Oxford Bulletin of Economics and Statistics* **79**, 495–512 (2017). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/obes.12163.

[29] Taylor, J. W. & Taylor, K. S. Combining probabilistic forecasts of COVID-19 mortality in the United States. *European Journal of Operational Research* (2021). URL https://www.sciencedirect.com/science/article/pii/S0377221721005609.

[30] Ray, E. L. *et al.* Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *International Journal of Forecasting* (2022). URL https://doi.org/10.1016/j.ijforecast.2022.06.005.

[31] Davies, N. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2021).

[32] Berndt, C., Endt, C. & Müller-Hansen, S. Die unsichtbare Welle. *Süddeutsche Zeitung* (2021). Published online, 5 February 2021, https://www.sueddeutsche.de/wissen/coronavirus-mutante-b117-daten-1.5197700.

[33] Fischer-Fels, J. Erste Hochrechnung zur Verbreitung der Coronamutationen. *Ärzteblatt* (2021). Published online, 3 February 2021, https://www.aerzteblatt.de/nachrichten/120768/Erste-Hochrechnung-zur-Verbreitung-der-Corona-Mutationen.

[34] Landesgesundheitsamt Baden Württemberg. Tagesbericht COVID-19, Montag 8.2.2021 (2021). Available at https://www.gesundheitsamt-bw.de/fileadmin/LGA/_DocumentLibraries/SiteCollectionDocuments/05_Service/Lagebericht COVID19/COVID_Lagebericht_LGA_210208.pdf.

[35] Robert Koch Institute. Bericht zu Virusvarianten von SARS-CoV-2 in Deutschland, insbesondere zur Variant of Concern (VOC) B.1.1.7, update 10 February 2021. Available at https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/Bericht_VOC_2021-02-10.pdf (2021).

[36] Reich, N., Tibshirani, R., Ray, E. & Rosenfeld, R. On the predictability of COVID-19. Blog post, International Institute of Forecasters, https://forecasters.org/blog/2021/09/28/on-the-predictability-of-covid-19/ (2021).

[37] Keyel, A. C. & Kilpatrick, A. M. Probabilistic evaluation of null models for West Nile Virus in the United States (2021). URL https://www.biorxiv.org/content/early/2021/07/26/2021.07.26.453866.

[38] Funk, S. *et al.* Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. *medRxiv* (2020). URL https://www.medrxiv.org/content/early/2020/11/13/2020.11.11.20220962.

[39] Ray, E. L. *et al.* Challenges in training ensembles to forecast COVID-19 cases and deaths in the United States. Blog post, International Institute of Forecasters, https://forecasters.org/blog/2021/04/09/challenges-in-training-ensembles-to-forecast-covid-19-cases-and-deaths-in-the-united-states/ (2021).

[40] Castro, M., Ares, S., Cuesta, J. & Manrubia, S. The turning point and end of an expanding epidemic cannot be precisely forecast. *Proceedings of the National Academy of Sciences* **117**, 26190–26196 (2020). URL https://doi.org/10.1073/pnas.2007868117.

[41] Wilke, C. O. & Bergstrom, C. T. Predicting an epidemic trajectory is difficult. *Proceedings of the National Academy of Sciences* **117**, 28549–28551 (2020). URL https://doi.org/10.1073/pnas.2020200117.

[42] Harvey, A. Time series modelling of epidemics: Leading indicators, control groups and policy assessment. *National Institute Economic Review* **257**, 83–100 (2021).

[43] McDonald, D. J. *et al.* Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction? *Proceedings of the National Academy of Sciences* **118**, e2111453118 (2021). URL https://www.pnas.org/doi/abs/10.1073/pnas.2111453118. https://www.pnas.org/doi/pdf/10.1073/pnas.2111453118.

[44] Fuhrmann, J. & Barbarossa, M. The significance of case detection ratios for predictions on the outcome of an epidemic - a message from mathematical modelers. *Archives of Public Health* **78**, article number 63 (2020).

[45] Berndt, C., Hametner, M., Kruse, B., Müller-Hansen, S. & Witzenberger, B. Ist die dritte Welle überstanden? *Süddeutsche Zeitung* (2021). Published online, 4 May 2020, https://www.sueddeutsche.de/gesundheit/corona-infektionen-trendwende-modellierungen-1.5284545.

[46] Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. & Gneiting, T. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science* **32**, 106–127 (2017).

[47] Coibion, O. & Gorodnichenko, Y. What can survey forecasts tell us about information rigidities? *Journal of Political Economy* **120**, 116–159 (2012).

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

[49] Adamik, B. *et al.* Mitigation and herd immunity strategy for COVID-19 is likely to fail. *medRxiv* (2020). URL https://doi.org/10.1101/2020.03.25.20043109.

[50] Li, M. L. *et al.* Forecasting COVID-19 and analyzing the effect of government interventions. *medRxiv* (2020). URL https://www.medrxiv.org/content/early/2020/06/24/2020.06.23.20138693.

[51] Barbarossa, M. V. *et al.* Modeling the spread of COVID-19 in Germany: Early assessment and possible scenarios. *PLOS ONE* **15**, e0238559 (2020).

[52] Karlen, D. Characterizing the spread of CoViD-19. *arXiv preprint arXiv:2007.07156* (2020). URL https://arxiv.org/abs/2007.07156.

[53] Kheifetz, Y., Kirsten, H. & Scholz, M. On the parametrization of epidemiologic models – lessons from modelling COVID-19 epidemic (2022). *Viruses* **14**, 1468.

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

[55] Abbott, S. *et al.* Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research* **5** (2020). URL https://doi.org/10.12688/wellcomeopenres.15842.3.

[56] Krymova, E. *et al.* Trend estimation and short-term forecasting of COVID-19 cases and deaths worldwide. *Proceedings of the National Academy of Sciences of the USA* **119**, e2112656119 (2022).

[57] Burgard, J. P., Heyder, S., Hotz, T. & Krueger, T. Regional estimates of reproduction numbers with application to COVID-19. *arXiv preprint arXiv:2108.13842* (2021). URL https://arxiv.org/abs/2108.13842.

[58] Castro, L., Fairchild, G., Michaud, I. & Osthus, D. COFFEE: COVID-19 forecasts using fast evaluations and estimation. *arXiv preprint arXiv:2110.01546* (2021). URL https://arxiv.org/abs/2110.01546.

[59] Bosse, N. I. *et al.* Comparing human and model-based forecasts of COVID-19 in Germany and Poland. *medRxiv* (2021). URL https://www.medrxiv.org/content/early/2021/12/05/2021.12.01.21266598.

[60] Bhatia, S. *et al.* Global predictions of short- to medium-term COVID-19 transmission trends : a retrospective assessment. *medRxiv* (2021). URL https://www.medrxiv.org/content/early/2021/07/22/2021.07.19.21260746.

[61] Bracher, J., Wolffram, D., & the German and Polish COVID-19 Forecast Hub Team. Codes underlying the analyses in Bracher, Wolffram et al: National and subnational short-term forecasting of COVID-19 in Germany and Poland during early 2021. Available online, https://zenodo.org/record/5639514#.Yv5fUmFBxH5, https://doi.org/10.5281/zenodo.5639514 last accessed on 18 August 2022. (2022).

[62] Zeileis, A. *et al.* colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. *Journal of Statistical Software* **96**(1), 1–49 (2020). URL https://doi.org/10.18637/jss.v096.i01.

[63] Lemon, J. Plotrix: a package in the red light district of R. *R News* **6**(4), 8–12 (2006). URL https://cran.r-project.org/doc/Rnews/Rnews_2006-4.pdf.

[64] Dahl, D.B. *et al.* xtable: Export Tables to LaTeX or HTML. R package version 1.8-4, URL https://cran.r-project.org/web/packages/xtable/. (2019).

[65] Zeileis, A. & Grothendieck, G. zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software* **14**(6), 1–27 (2005). URL https://doi.org/10.18637/jss.v014.i06.

Acknowledgements

J. Bracher, M. Schienle and T. Gneiting acknowledge support from the Helmholtz Foundation via the SIMCARD Information and Data Science Pilot Project. J. Bracher and M. Schienle were moreover supported by the German Ministry of Reserach and Education via the project RESPINOW. T. Gneiting and D. Wolffram are grateful for support by the Klaus Tschira Foundation. D. Wolffram's contribution was moreover supported by the Helmholtz Association under the joint research school HIDSS4Health – Helmholtz Information and Data Science School for Health as well as the German Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments. N.I. Bosse was supported by the Health Protection Research Unit (grant code NIHR200908). S. Funk and S. Abbott were supperted by the Wellcome Trust (210758/Z/18/Z). The itwm-dSEIR forecasting team (J. Fiedler, N. Leithäuser, J. Mohring) was supported by the Ministry of Health and Science of Rhineland Palatinate and the Fraunhofer Anti-Corona Program. The LANL-GrowthRate forecasting team (L. Castro, G. Fairchild, I.J Michaud) was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20200700ER. S. Bhatia acknowledges funding from the Wellcome Trust (219415). Work on the ICM UW epidemiological model (J.M Nowosielski, M. Radwan, F. Rakowski) was supported by the Polish Minister of Science and Higher Education grant 51/WFSN/2020 given to the University of Warsaw. Development of the IMISE-SECIR model (Y. Kheifetz, H. Kirsten, M. Scholz) was funded in the framework of the project SaxoCOV (Saxonian COVID-19 Research Consortium). SaxoCOV was co-financed with tax funds on the basis of the budget passed by the Saxon state parliament. Model presentation was funded by the NFDI4Health Task Force COVID-19 (www.nfdi4health.de/task-force-covid-19-2) within DFG project LO-342/17-1. Furthermore, modeling of this group is funded by the Federal Ministry of Education and Research Germany (BMBF) within project PROGNOSIS (FKZ 031L0296A).

We thank Dean Karlen for contributing forecasts and the Institute for Health Metrics and Evaluation, University of Washington, for making forecasts publicly available under a free license. We are moreover grateful for support and advice from the organizing team of the US COVID-19 Forecast Hub.

The content of this manuscript 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, DW, TG and MSe 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, MSe and members of

various teams. SA, MVB, DB, SB, MB, NIB, JPB, LC, GF, JFr, JFn, SF, AG, KG, SH, TH, YK, HK, TK, EK, NL, MLL, JHM, BM, IJM, JMe, JMg, PN, JMN, TO, MR, FR, MSz, SS, and AS contributed forecasts (see list of contributors by team). JB, TG and MSe wrote the manuscript. All teams and members of the coordinating team provided feedback on the manuscript and descriptions of the respective models.

List of contributors by team

CovidAnalytics-DELPHI: Michael Lingzhi Li (Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA), Dimitris Bertsimas, Saksham Soni (both Sloan School of Management, Massachusetts Institute of Technology, Cambridge, 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 (Institute of Applied Mathematics, University of Heidelberg, Heidelberg, Germany), Jan H. Meinke (Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany)

SDSC_ISG-TrendModel: Antoine Flahault, Elisa Manetti, Kristen Namigai (all 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: Filip Dreger, 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)

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), Pierre Nouvellet (School of Life Sciences, University of Sussex, Brighton, UK)

itwm-dSEIR: Michael Burger, Robert Feßler, Jochen Fiedler, Michael Helmling, Karl-Heinz Küfer, Neele Leithäuser, Jan Mohring, Johanna Schneider, Anita Schöbel, Michael Speckert, Raimund Wegener, Jarosław Wlazło (all Fraunhofer Institute for Industrial Mathematics, Kaiserslautern, Germany)

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, Ilmenau, Germany)

LANL-GrowthRate: Dave A. Osthus, Isaac J. Michaud (both Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, USA), Lauren Castro, Geoffrey Fairchild (both Information Systems and Modeling, Los Alamos National Laboratory, Los Alamos, USA)

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, Radosław Idzikowski, Tyll Krueger, Tomasz Ożański, Ewaryst Rafajłłowicz, Ewa Skubalska-Rafajłowicz, Wojciech Rafajłłowicz (all Wroclaw University of Science and Technology, Poland), Barbara Pabjan (Institute of Sociology,, University of Wroclaw, 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)

USC-SIkJalpha: Ajitesh Srivastava, Frost Tianjian Xu (both University of Southern California, Los Angeles, USA)

Supplementary Materials for Bracher et al (2021): National and subnational short-term forecasting of COVID-19 in Germany and Poland during early 2021

Johannes Bracher[1,2,*], Daniel Wolffram[1,2,3], Jannik Deuschel[1], Konstantin Görgen[1], Jakob L. Ketterer[1], Alexander Ullrich[4], Sam Abbott[5], Maria V. Barbarossa[6], Dimitris Bertsimas[7], Sangeeta Bhatia[8], Marcin Bodych[9], Nikos I. Bosse[5], Jan Pablo Burgard[10], Lauren Castro[11], Geoffrey Fairchild[11], Jochen Fiedler[12], Jan Fuhrmann[13], Sebastian Funk[5], Anna Gambin[14], Krzysztof Gogolewski[14], Stefan Heyder[15], Thomas Hotz[15], Yuri Kheifetz[16], Holger Kirsten[16], Tyll Krueger[9], Elena Krymova[17], Neele Leithäuser[12], Michael L. Li[18], Jan H. Meinke[19], Błażej Miasojedow[14], Isaac J. Michaud[20], Jan Mohring[12], Pierre Nouvellet[21], Jedrzej M. Nowosielski[22], Tomasz Ozanski[9], Maciej Radwan[22], Franciszek Rakowski[22], Markus Scholz[16], Saksham Soni[18], Ajitesh Srivastava[23], Tilmann Gneiting[2,24], Melanie Schienle[1,*]

August 25, 2022

[1]Chair of Statistics and Econometrics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2]Computational Statistics Group, Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany
[3]HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[4]Robert Koch Institute (RKI), Berlin, Germany
[5]London School of Hygiene and Tropical Medicine, London, UK
[6]Frankfurt Institute for Advanced Studies, Frankfurt, Germany
[7]Sloan School of Management, Massachusetts Institute of Technology, Cambridge MA, USA
[8]MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), Imperial College London, London, UK
[9]Wroclaw University of Science and Technology, Wroclaw, Poland
[10]Economic and Social Statistics Department, University of Trier, Trier, Germany
[11]Information Systems and Modeling, Los Alamos National Laboratory, Los Alamos, USA
[12]Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany
[13]Institute for Applied Mathematics, University of Heidelberg, Heidelberg, Germany
[14]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland
[15]Institute of Mathematics, Technische Universität Ilmenau, Ilmenau, Germany
[16]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany
[17]Swiss Data Science Center, ETH Zürich and EPF Lausanne, Zürich, Switzerland
[18]Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA
[19]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany
[20]Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, USA
[21]School of Life Sciences, University of Sussex, Brighton, UK
[22]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland
[23]Ming Hsieh Department of Computer and Electrical Engineering, University of Southern California, Los Angeles, USA
[24]Institute for Stochastics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

S1 Detailed description of new models

We only provide detailed descriptions of models which were added to our project for the second evaluation period. Descriptions for the other models can be found in Supplementary Note 3 of Bracher et al (2021). A more detailed documentation of the `LeipzigIMISE-SECIR` and `SDSC_ISG-TrendModel` models which had not been available at the appearance of Bracher (2021) can be found in Kheifetz et al (2021) and Krymova et al (2021), respectively.

itwm-dSEIR Fraunhofer-ITWM's predictions are based on a cohort model that groups people according to four age groups and according to the status infected, detected and since 19 April successfully vaccinated (i.e., this extension was added after the evaluation period). The dynamics of the epidemic are described by integral equations, assuming an infectious period with fixed onset, end and infectivity. The most important parameters are contact rates between age groups, detection rates and times, and death rates and times, which are adjusted to the historical data of the RKI. For forecasts, the simulation is continued with the parameters determined for the last week. In principle, the forecast quality could be improved by anticipating the effects of events such as the end of public holidays on contact and detection rates. However, this is not yet done in the automatic submissions. All calculations use automatic differentiation. This speeds up parameter adjustment and allows for error estimates. The latter are determined by comparing counted and simulated cases and by matching the empirical standard deviations with the standard deviations predicted by the calculated sensitivities. The model is described in detail in https://www.itwm.fraunhofer.de/de/presse-publikationen/presseinformationen/2021/2021-06-22_Dritte_Welle_Starker-Effekt-von-Schnelltests-an-Schulen.html.

Karlen-pypm The python Population Modeller (pyPM, Karlen 2020) is a mechanistic modeling framework to describe viral spread via discrete-time difference equations. In a pyPM model, different population objects are connected by a list of directional connector objects. The adjustable parameters of the model are stored in parameter objects. The core of the model consists of a model of the infection cycle involving the susceptible, infected (but not yet contagious) and contagious parts of the population. The contagious population is modelled in more detail by introducing symptomatic, test-positive, hospitalized (normal ward and ICU) and deceased populations. The model takes time series of cases, deaths and intensive care occupancy as data inputs. Forecasts are generated at the regional level (German states) first and subsequently aggregated to the national level. Starting from 1 March 2021, the model was stratified into spread of the wild type of the virus and the B.1.1.7 variant, and integrated genetic sequencing data on their respective importance.

S2 Additional forecast visualizations

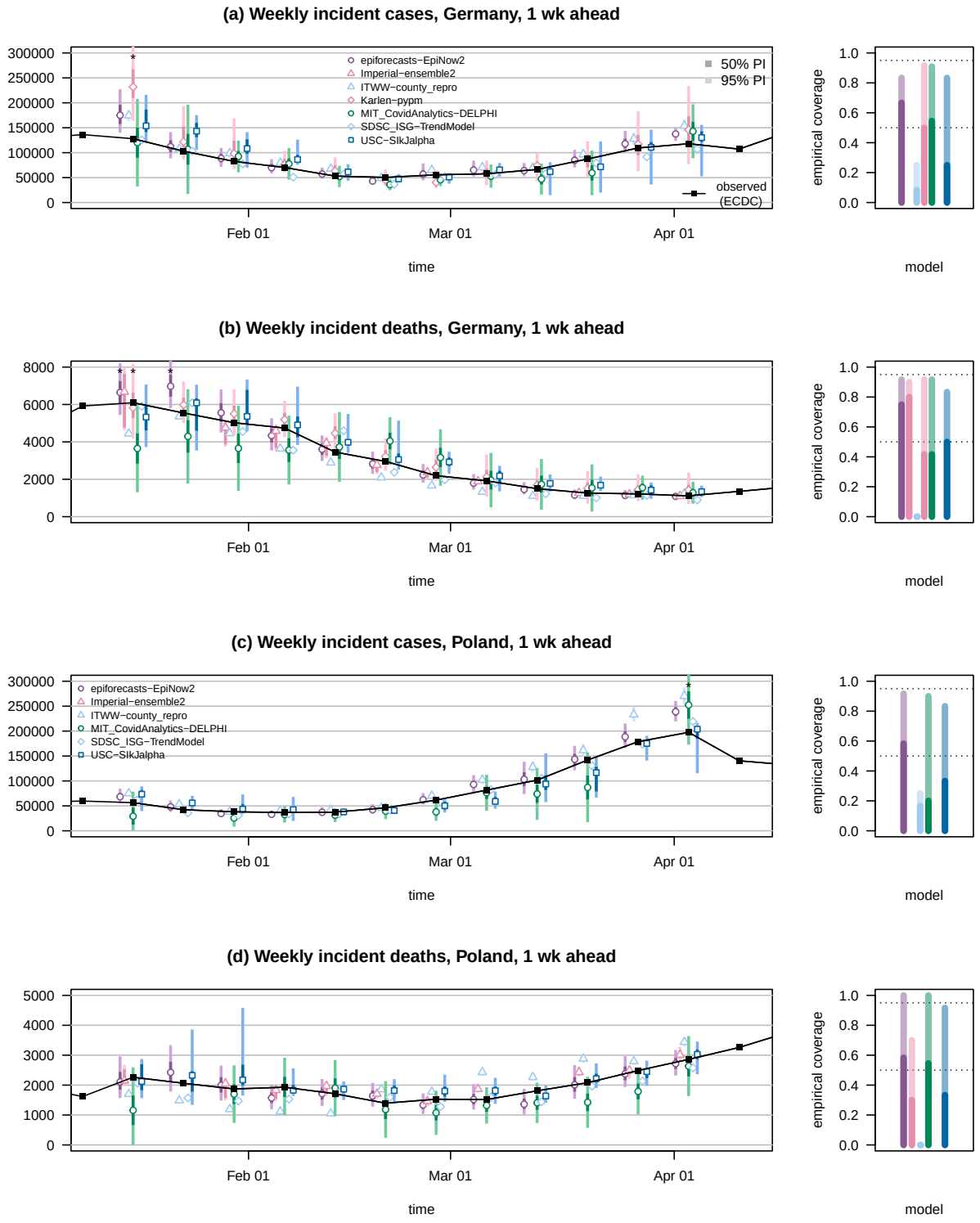


Figure S1: One-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. Asterisks mark prediction intervals exceeding the upper plot limit. The figure shows forecasts from models not displayed in Figure 2. Colored points represent predictive medians, dark and light bars show 50% and 95% prediction intervals, respectively. Asterisks mark prediction intervals exceeding the upper plot limit.

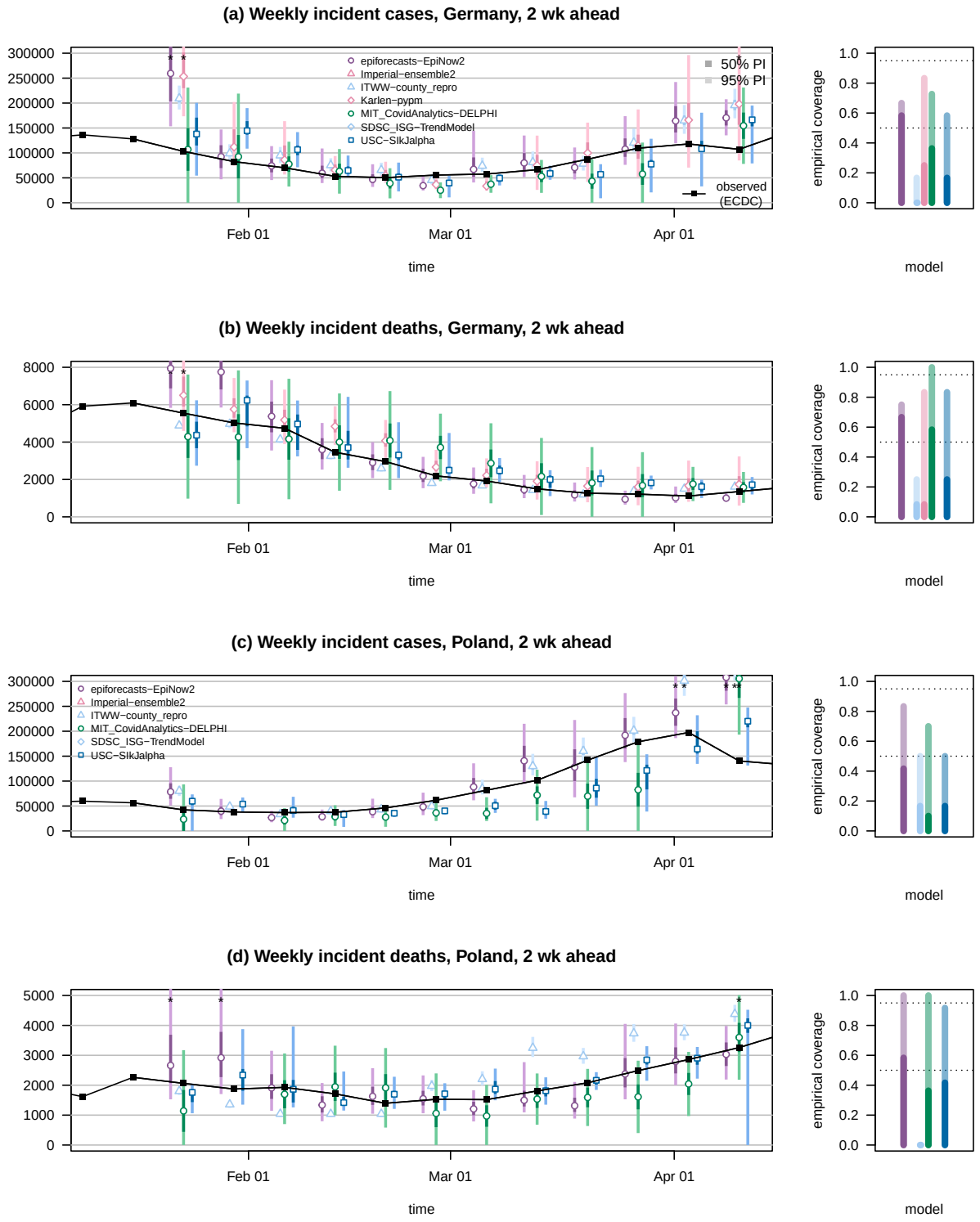


Figure S2: Two-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 3. Colored points represent predictive medians, dark and light bars show 50% and 95% prediction intervals, respectively. Asterisks mark prediction intervals exceeding the upper plot limit.

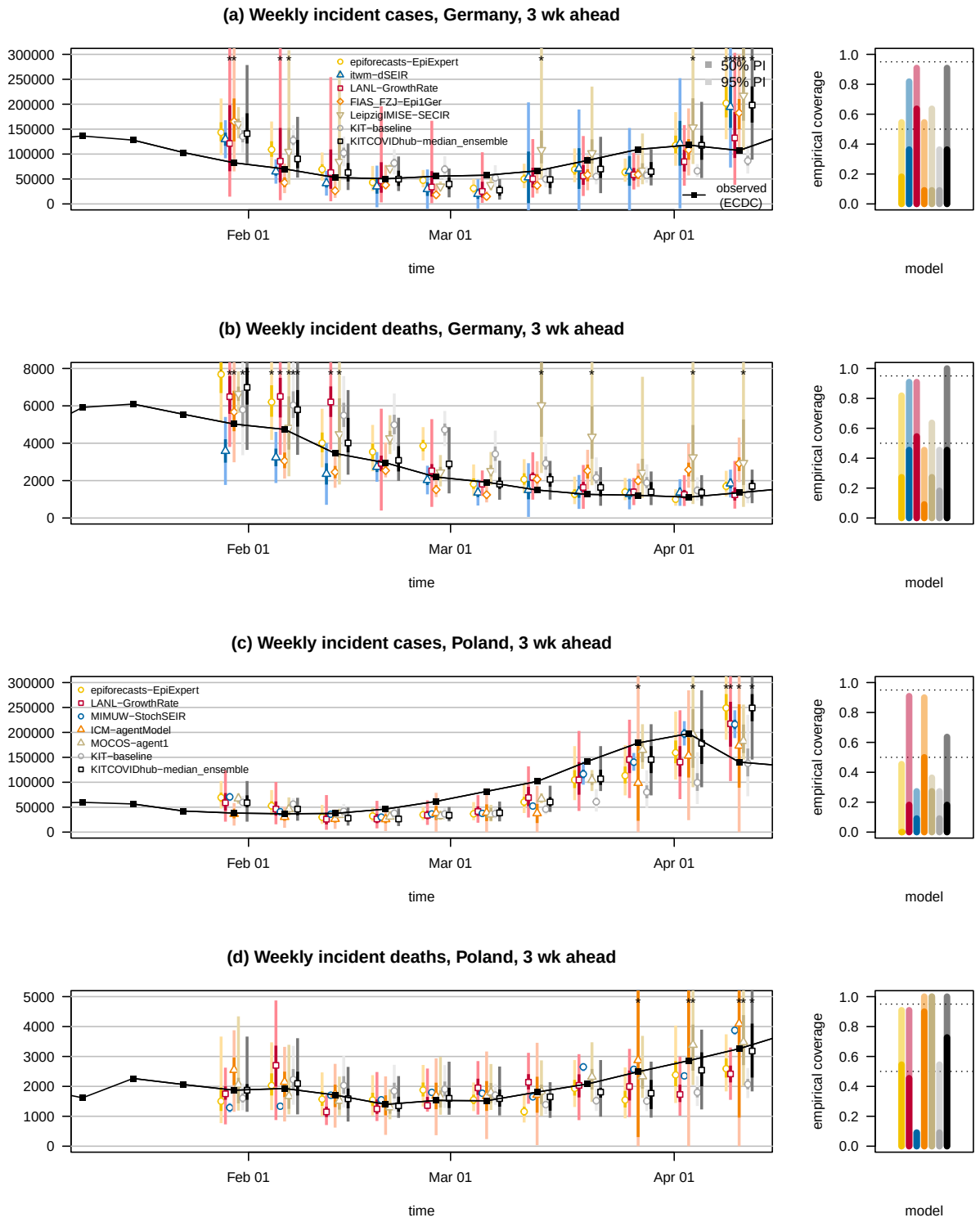


Figure S3: Three-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 3. Colored points represent predictive medians, dark and light bars show 50% and 95% prediction intervals, respectively. Asterisks mark prediction intervals exceeding the upper plot limit.

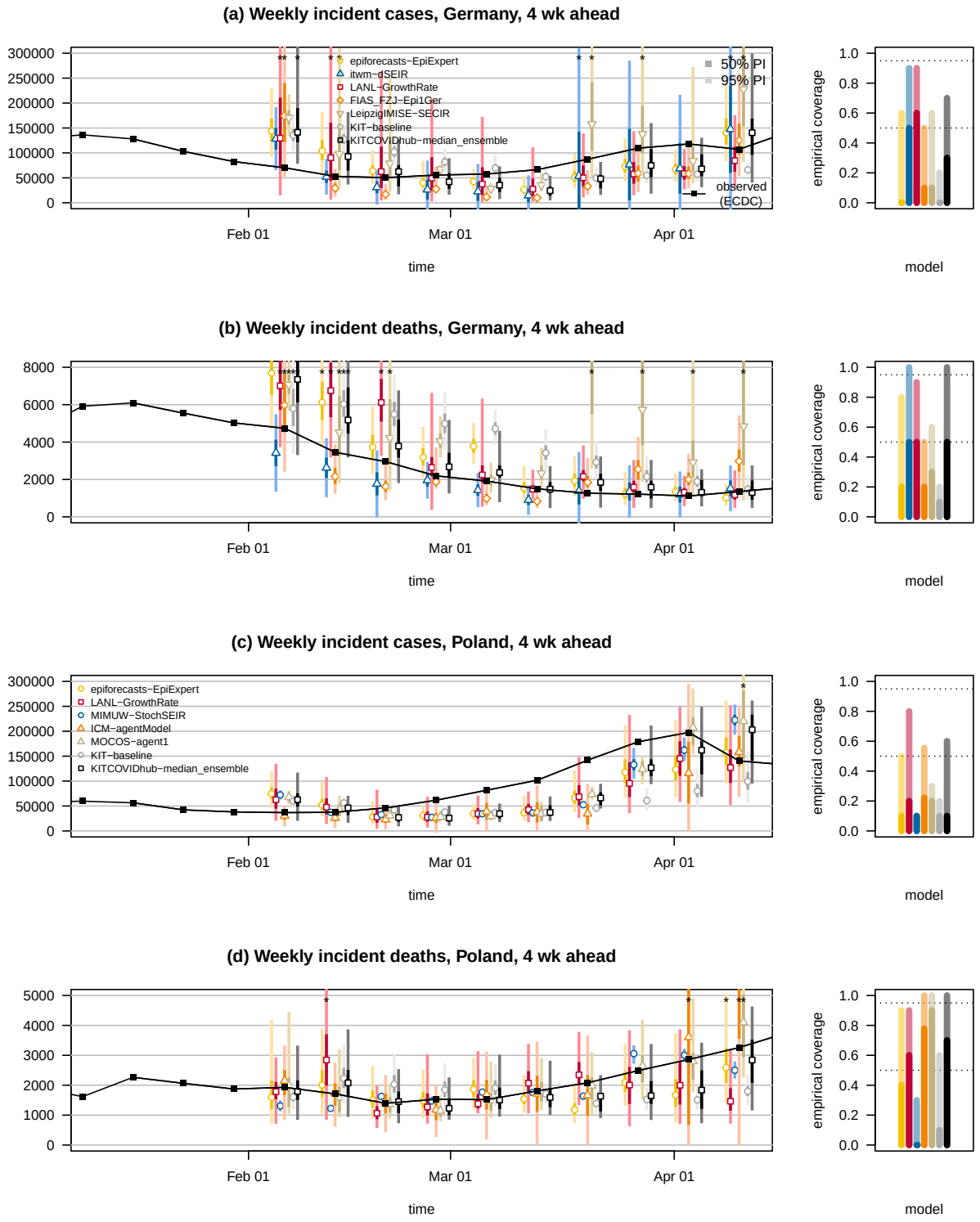


Figure S4: Four-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 3. Colored points represent predictive medians, dark and light bars show 50% and 95% prediction intervals, respectively. Asterisks mark prediction intervals exceeding the upper plot limit.

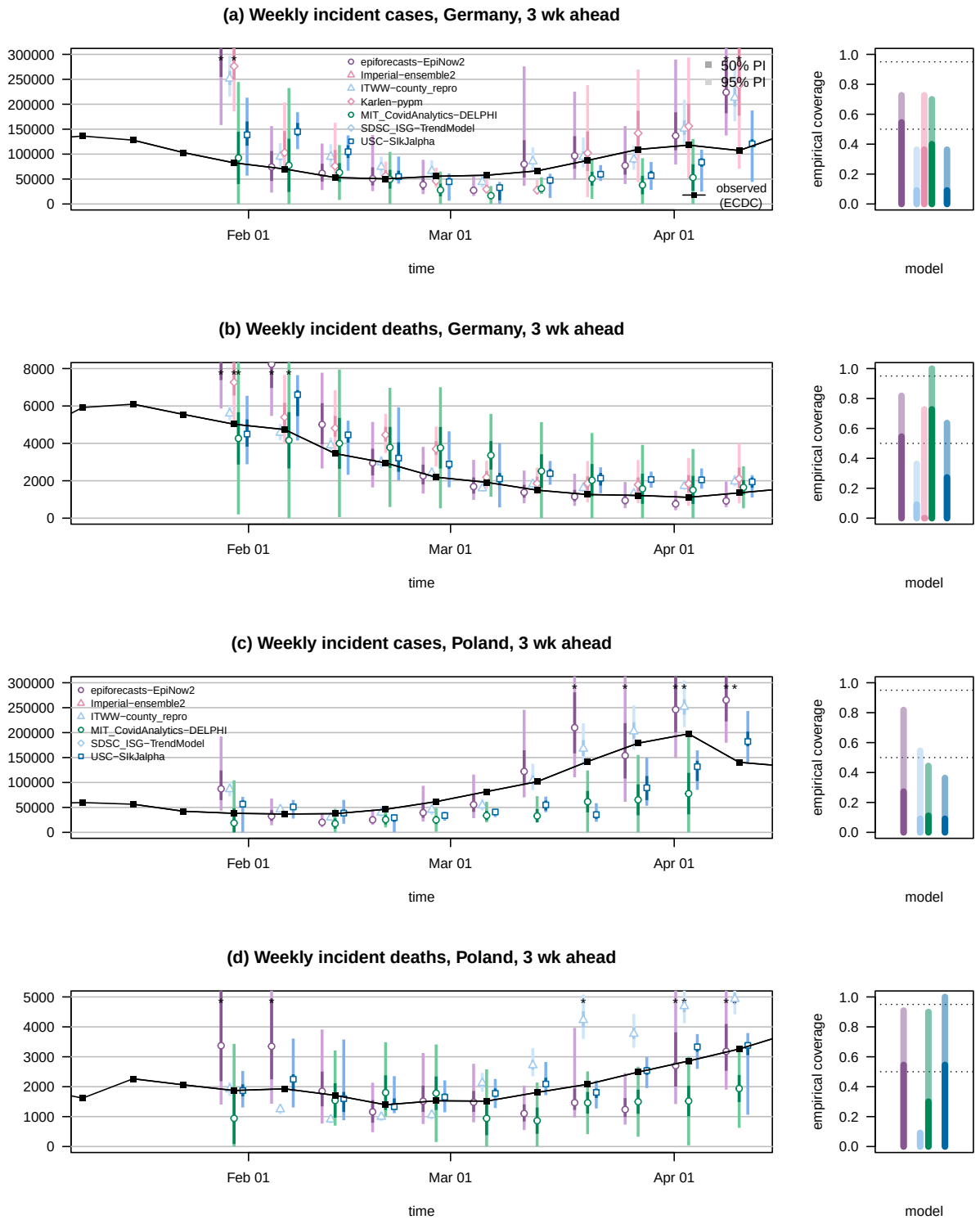


Figure S5: Three-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure S1. Colored points represent predictive medians, dark and light bars show 50% and 95% prediction intervals, respectively. Asterisks mark prediction intervals exceeding the upper plot limit.

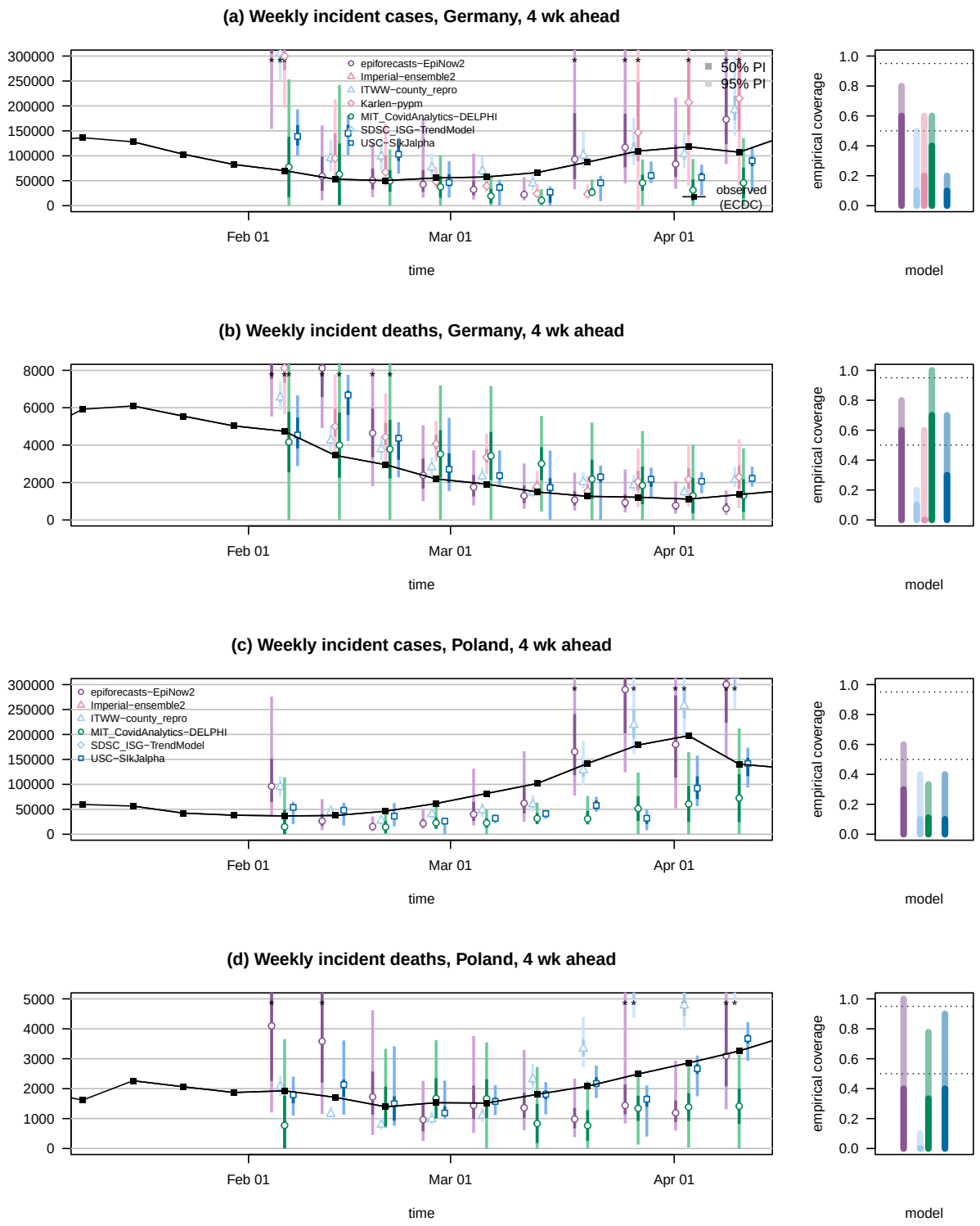


Figure S6: Four-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure S1. Colored points represent predictive medians, dark and light bars show 50% and 95% prediction intervals, respectively. Asterisks mark prediction intervals exceeding the upper plot limit.

S3 Decomposition of average WIS values

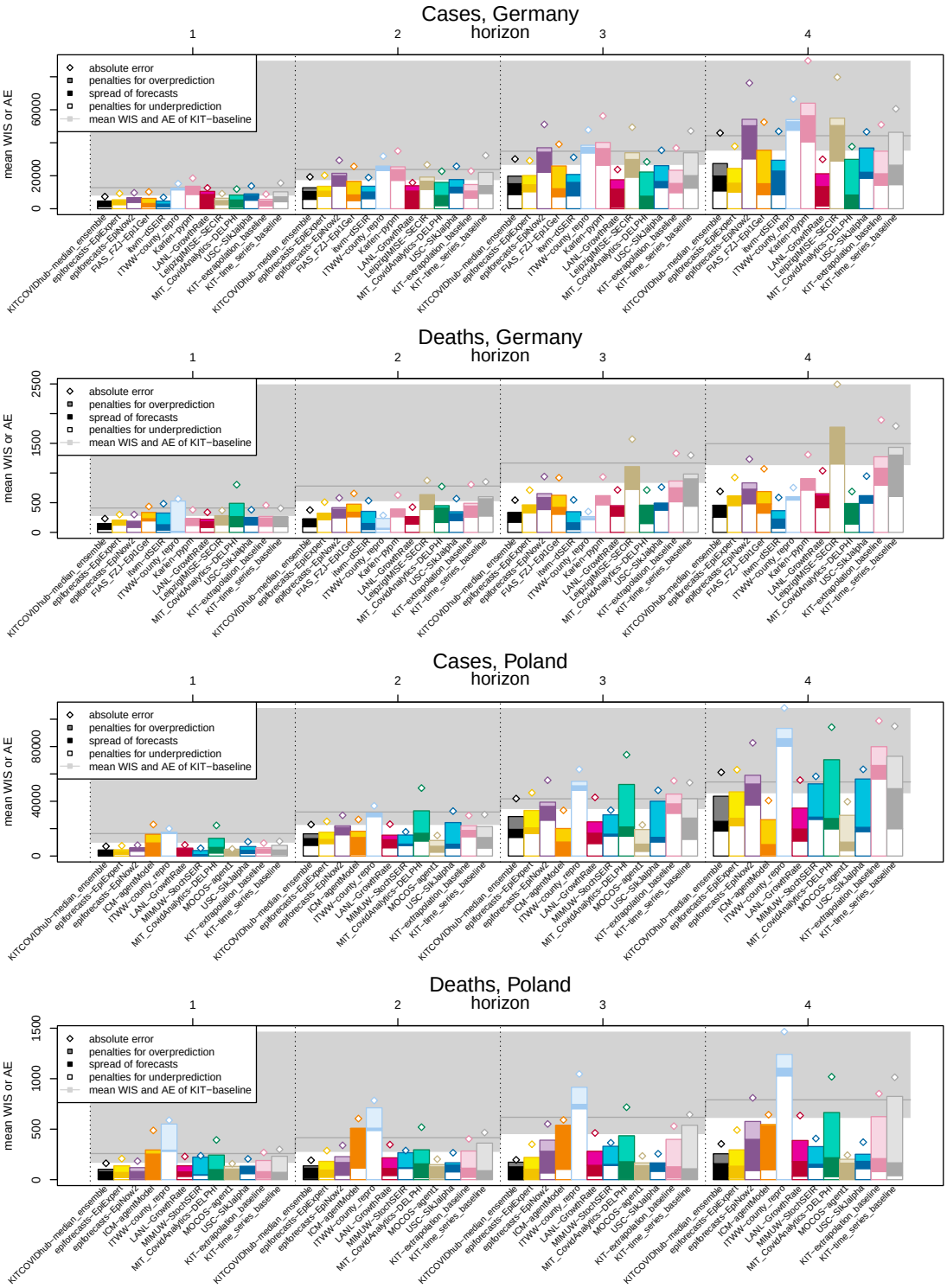


Figure S7: Average weighted interval score and absolute error achieved by models across countries, targets and forecast horizons. The grey area represents the performance of the baseline model `KIT-baseline`. WIS values are decomposed into components for forecast spread, overprediction and underprediction.

S4 Individual WIS values

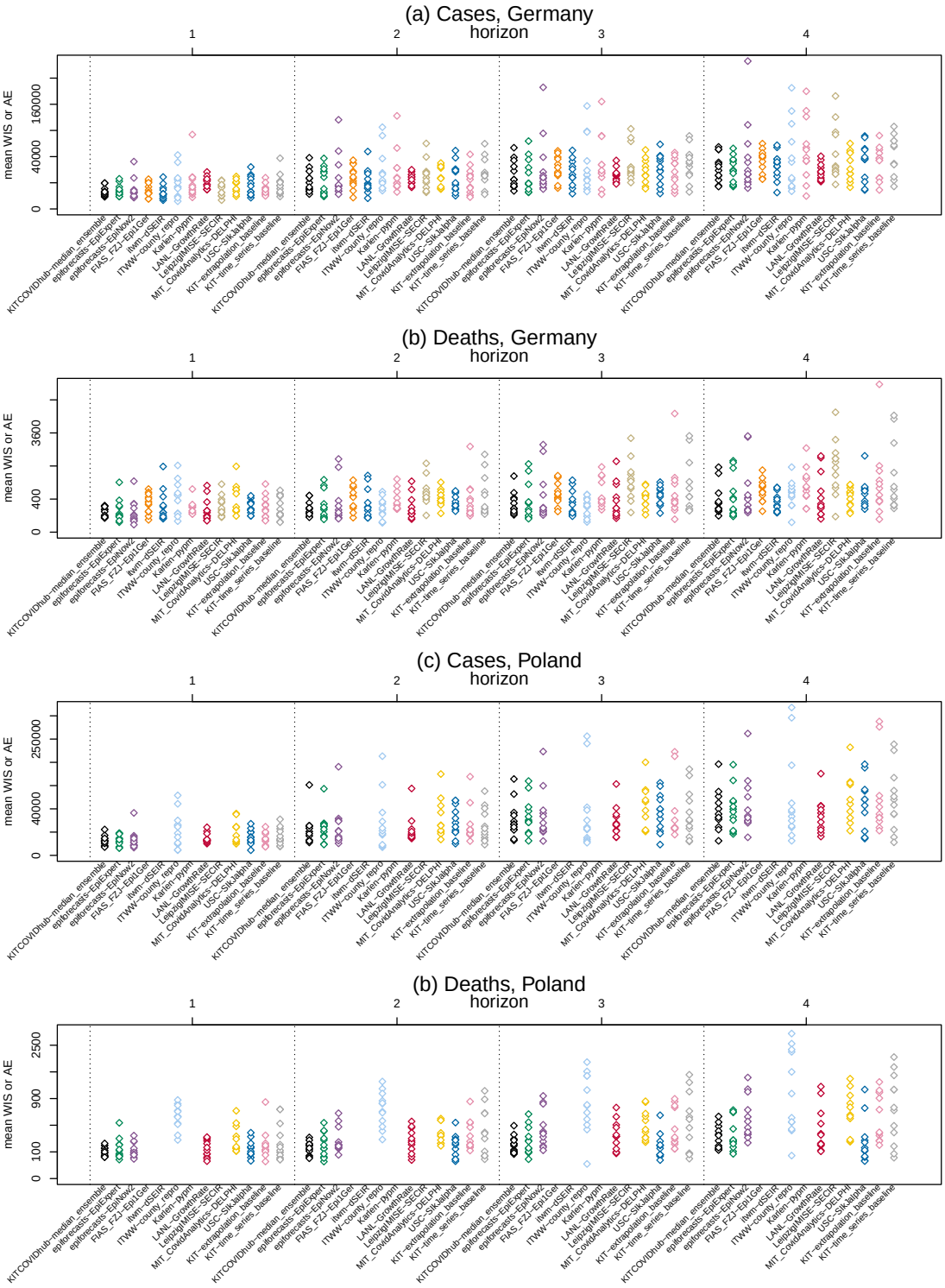


Figure S8: Individual weighted interval scores achieved by models across countries, targets and forecast horizons. Each dot represents one score achieved over the 12-week evaluation period.

S5 Weights in inverse-WIS ensembles

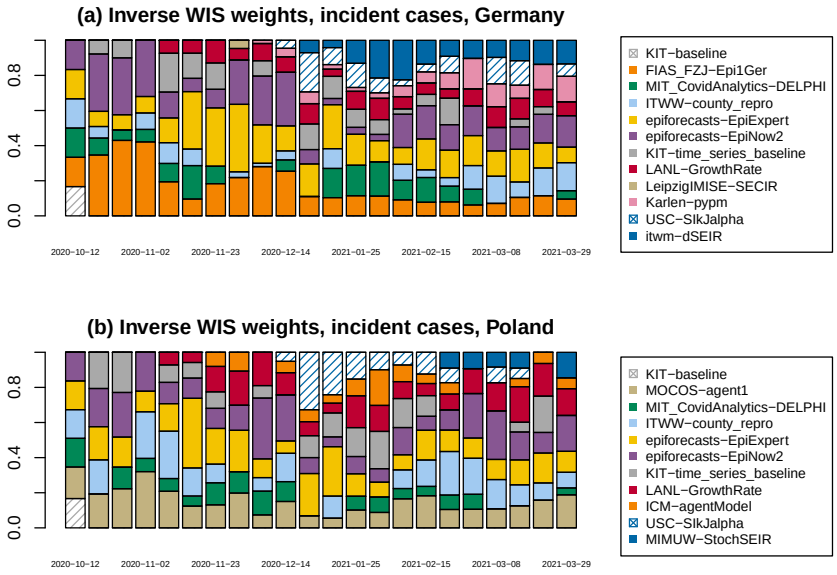


Figure S9: Weights in `KITCOVIDhub-inverse_wis_ensemble` for incident cases in Germany and Poland, October 2020–March 2021 (i.e., combined for the study periods of Bracher et al (2021) and the present manuscript).

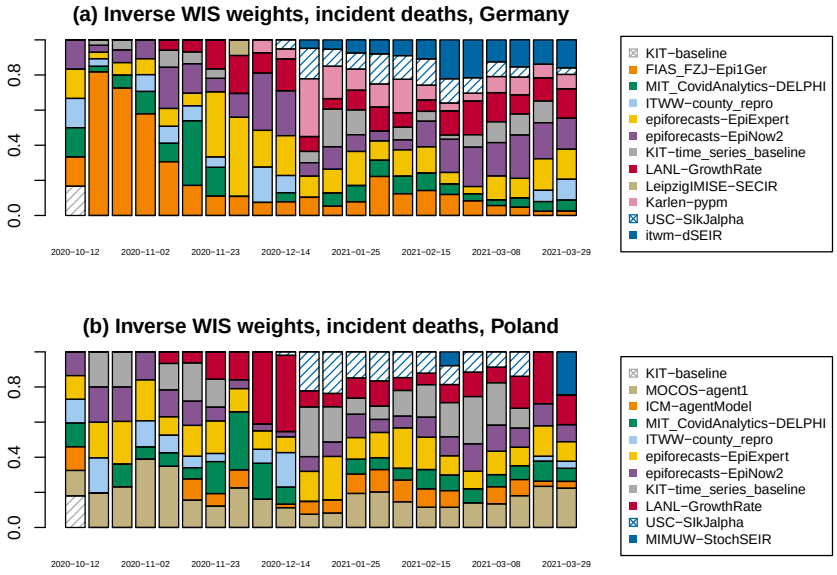


Figure S10: Weights in `KITCOVIDhub-inverse_wis_ensemble` for incident deaths in Germany and Poland, October 2020–March 2021 (i.e., combined for the study periods of Bracher et al (2021) and the present manuscript).

S6 Visualization of behaviour at turning points

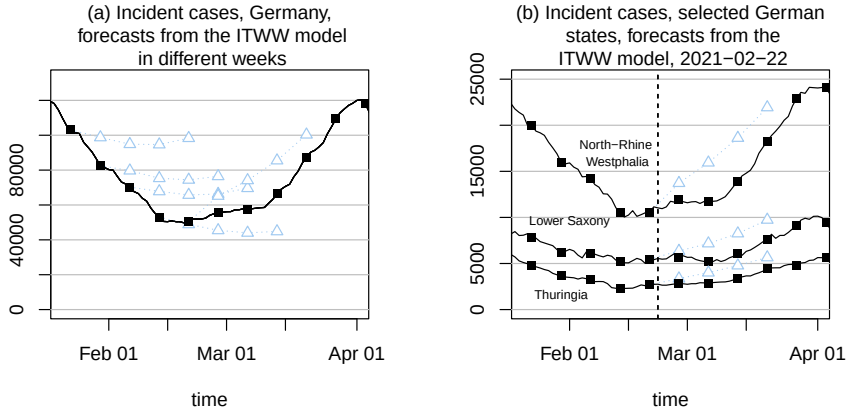


Figure S11: **a** Forecasts of cases in Germany by the `ITWW-county_repro` model, 25 January to 22 February 2021. **b** Forecasts for cases in selected German states by the `ITWW-county_repro` model, 22 February 2021.

S7 Additional summary tables on forecast evaluation

Table S1: Forecast evaluation for Germany and Poland in terms of relative AE and WIS, 1–4 weeks ahead. The relative values are obtained by dividing the mean absolute error or WIS of a given model by the respective value achieved by the baseline model. Values below 1 indicate better performance than the baseline, values above worse performance.

Germany

	1 wk case		2 wk case		3 wk case		4 wk case		1 wk death		2 wk death		3 wk death		4 wk death	
Model	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS
epiforecasts-EpiExpert	0.73	0.68	0.85	0.79	0.83	0.76	0.86	0.69	0.73	0.74	0.65	0.61	0.61	0.56	0.62	0.54
epiforecasts-EpiNow2	0.76	0.84	1.23	1.24	1.47	1.39	1.72	1.54	0.73	0.68	0.74	0.79	0.80	0.78	0.83	0.74
FIAS_FZJ-Epi1Ger	0.80	0.79	1.08	0.96	1.12	0.98	1.19	1.01	1.06	1.21	0.84	0.91	0.79	0.75	0.72	0.61
IHME-CurveFit									1.26		0.84		0.72		0.65	
Imperial-ensemble2									0.47	0.49						
itwm-dSEIR	0.54	0.58	0.80	0.79	0.89	0.78	1.06	0.83	1.18	1.18	0.68	0.67	0.47	0.42	0.39	0.32
ITWW-county_repro	1.19	1.56	1.34	1.49	1.37	1.45	1.50	1.53	1.37	1.91	0.37	0.45	0.30	0.32	0.50	0.53
Karlen-pypm	1.45	1.71	1.47	1.46	1.61	1.51	2.03	1.82	0.93	0.84	0.80	0.75	0.80	0.74	0.88	0.80
LANL-GrowthRate	0.99	1.33	0.66	0.80	0.68	0.66	0.68	0.60	0.82	0.80	0.54	0.51	0.61	0.54	0.69	0.57
LeipzigIMISE-SECIR	0.72	0.80	1.12	1.11	1.42	1.28	1.80	1.56	0.90	1.01	1.12	1.21	1.34	1.33	1.67	1.57
MIT_CovidAnalytics-DELPHI	0.93	1.04	0.96	0.92	0.81	0.84	0.85	0.85	1.95	1.77	0.99	0.86	0.61	0.55	0.46	0.43
SDSC_ISG-TrendModel	0.62								1.06							
USC-SIkJalpha	1.08	1.13	1.08	1.02	1.02	0.98	1.05	1.05	0.93	0.92	0.73	0.66	0.65	0.59	0.63	0.54
KIT-baseline	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
KIT-extrapolation_baseline	0.69	0.72	0.96	0.85	1.06	0.88	1.15	0.99	1.11	0.97	1.03	0.93	1.14	1.04	1.27	1.12
KIT-time_series_baseline	1.22	1.29	1.36	1.27	1.35	1.28	1.37	1.32	0.99	0.95	1.09	1.15	1.11	1.18	1.20	1.26
KITCOVIDhub-inverse_wis_ensemble	0.67	0.67	0.92	0.80	0.98	0.81	1.11	0.86	0.53	0.54	0.39	0.39	0.42	0.38	0.45	0.40
KITCOVIDhub-mean_ensemble	0.66	0.66	0.92	0.79	0.97	0.80	1.10	0.84	0.53	0.55	0.44	0.42	0.44	0.38	0.48	0.38
KITCOVIDhub-median_ensemble	0.58	0.59	0.81	0.73	0.86	0.74	1.04	0.78	0.56	0.54	0.48	0.43	0.46	0.40	0.46	0.40

Poland

	1 wk case		2 wk case		3 wk case		4 wk case		1 wk death		2 wk death		3 wk death		4 wk death	
Model	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS	relAE	relWIS
epiforecasts-EpiExpert	0.46	0.47	0.79	0.77	1.11	0.97	1.16	1.02	0.81	0.82	0.69	0.66	0.57	0.49	0.62	0.48
epiforecasts-EpiNow2	0.48	0.61	0.92	0.97	1.32	1.15	1.53	1.29	0.71	0.71	0.82	0.83	0.89	0.87	1.02	0.94
ICM-agentModel	1.40	1.63	0.83	0.80	0.80	0.59	0.75	0.58	1.89	1.77	1.46	1.84	0.96	1.19	0.81	0.89
IHME-CurveFit									1.45		1.25		1.03		0.64	
Imperial-ensemble2									0.73	0.83						
ITWW-county_repro	1.22	1.78	1.14	1.38	1.51	1.59	2.00	2.04	2.28	3.31	1.89	2.59	1.69	2.03	1.85	2.03
LANL-GrowthRate	0.50	0.59	0.72	0.67	1.03	0.73	1.03	0.76	0.89	0.82	0.83	0.79	0.75	0.62	0.80	0.63
MIMUW-StochSEIR	0.35	0.41	0.55	0.68	0.80	0.88	1.08	1.15	0.92	1.34	0.69	0.97	0.59	0.73	0.52	0.54
MIT_CovidAnalytics-DELPHI	1.36	1.33	1.54	1.45	1.77	1.52	1.74	1.54	1.52	1.46	1.25	1.08	1.16	0.97	1.29	1.09
MOCOS-agent1	0.32	0.51	0.47	0.50	0.54	0.56	0.73	0.65	0.61	0.79	0.49	0.54	0.38	0.37	0.31	0.33
SDSC_ISG-TrendModel	0.39								1.02							
USC-SIkJalpha	0.63	0.71	1.02	1.08	1.15	1.17	1.17	1.23	0.80	0.80	0.64	0.61	0.41	0.37	0.47	0.41
KIT-baseline	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
KIT-extrapolation_baseline	0.58	0.62	0.92	0.98	1.32	1.32	1.82	1.74	1.04	1.14	0.97	1.03	0.86	0.89	1.08	1.02
KIT-time_series_baseline	0.66	0.80	0.94	0.95	1.28	1.22	1.75	1.59	1.16	1.39	1.12	1.32	1.04	1.20	1.28	1.35
KITCOVIDhub-inverse_wis_ensemble	0.45	0.48	0.73	0.69	0.92	0.79	1.10	0.93	0.58	0.67	0.47	0.52	0.39	0.39	0.52	0.43
KITCOVIDhub-mean_ensemble	0.42	0.49	0.74	0.69	0.99	0.80	1.16	0.96	0.55	0.69	0.42	0.55	0.27	0.41	0.45	0.43
KITCOVIDhub-median_ensemble	0.43	0.45	0.72	0.72	1.00	0.84	1.13	0.95	0.63	0.62	0.46	0.50	0.32	0.38	0.45	0.42

Table S2: Forecast evaluation for Germany and Poland, 3 and 4 weeks ahead (incidence scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	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	29,111	20,140	0.25	0.58	37,879	24,429	0.08	0.67	712	463	0.25	0.83	925	614	0.25	0.83
epiforecasts-EpiNow2	51,095	36,966	0.50	0.75	76,218	54,216	0.50	0.67	940	653	0.50	0.75	1,234	834	0.50	0.75
FIAS_FZJ-Epi1Ger	39,102	25,939	0.08	0.58	52,516	35,472	0.08	0.58	920	625	0.08	0.50	1,071	688	0.17	0.50
IHME-CurveFit									839				967			
Imperial-ensemble2																
itwm-dSEIR	31,177	20,772	0.33	0.83	46,895	29,388	0.42	0.92	547	348	0.42	0.92	588	366	0.42	1.00
ITWW-county_repro	47,796	38,477	0.08	0.33	66,520	54,070	0.08	0.42	350	266	0.08	0.33	753	599	0.08	0.17
Karlen-pypm	56,242	40,195	0.33	0.75	89,745	63,976	0.17	0.67	933	618	0.00	0.75	1,309	905	0.00	0.67
LANL-GrowthRate	23,760	17,577	0.67	0.92	29,959	21,224	0.67	0.92	715	450	0.50	0.92	1,037	651	0.50	0.92
LeipzigIMISE-SECIR	49,422	33,952	0.08	0.67	79,741	54,914	0.08	0.67	1,570	1,107	0.25	0.67	2,493	1,773	0.25	0.67
MIT_CovidAnalytics-DELPHI	*28,404	*22,268	0.45	0.73	*37,661	*29,934	0.45	0.64	713	459	0.75	1.00	688	486	0.75	1.00
SDSC_ISG-TrendModel																
USC-SIkJalpha	35,434	26,133	0.08	0.42	46,612	36,819	0.17	0.25	762	492	0.25	0.67	947	617	0.25	0.67
KIT-baseline	34,871	26,554	0.08	0.42	44,270	35,228	0.00	0.25	1,168	834	0.17	0.50	1,494	1,132	0.08	0.33
KIT-extrapolation_baseline	36,817	23,333	0.17	0.83	50,965	34,906	0.17	0.50	1,331	865	0.25	0.75	1,893	1,273	0.17	0.75
KIT-time_series_baseline	47,164	33,930	0.17	0.58	60,577	46,342	0.33	0.58	1,299	983	0.50	0.75	1,791	1,432	0.42	0.83
KITCOVIDhub-inverse_wis_ensemble	34,038	21,456	0.33	0.92	48,943	30,201	0.42	0.83	494	320	0.75	1.00	678	450	0.58	1.00
KITCOVIDhub-mean_ensemble	33,754	21,288	0.25	0.83	48,688	29,683	0.33	0.83	518	318	0.50	1.00	712	435	0.50	1.00
KITCOVIDhub-median_ensemble	30,154	19,767	0.33	0.92	45,928	27,397	0.25	0.75	543	332	0.42	1.00	690	456	0.42	1.00

Poland

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	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	46,220	33,301	0.00	0.42	63,055	46,882	0.08	0.42	350	220	0.58	0.92	491	295	0.42	0.92
epiforecasts-EpiNow2	55,368	39,353	0.25	0.75	82,709	59,002	0.25	0.50	552	391	0.58	0.92	810	576	0.50	1.00
ICM-agentModel	*33,401	*20,103	0.55	0.91	* 40,524	*26,560	0.36	0.64	* 592	*536	0.91	1.00	* 643	* 546	0.82	1.00
IHME-CurveFit									637				505			
Imperial-ensemble2																
ITWW-county_repro	63,327	54,555	0.08	0.50	108,188	93,286	0.08	0.33	1,047	915	0.08	0.08	1,466	1,241	0.00	0.08
LANL-GrowthRate	42,864	24,958	0.17	0.83	55,486	35,013	0.17	0.75	463	281	0.42	0.92	635	388	0.58	0.83
MIMUW-StochSEIR	33,532	30,118	0.08	0.25	58,203	52,730	0.08	0.08	365	331	0.08	0.08	408	327	0.08	0.33
MIT_CovidAnalytics-DELPHI	*74,004	*52,292	0.10	0.40	* 94,153	*70,385	0.10	0.30	* 717	*435	0.36	0.91	*1,019	* 665	0.27	0.73
MOCOS-agent1	22,746	19,350	0.33	0.42	39,711	29,751	0.25	0.42	235	166	0.92	1.00	243	202	0.92	1.00
SDSC_ISG-TrendModel																
USC-SIkJalpha	48,043	40,067	0.08	0.33	63,352	56,280	0.08	0.33	256	166	0.50	1.00	371	252	0.42	0.92
KIT-baseline	41,804	34,355	0.08	0.33	54,127	45,833	0.17	0.25	618	451	0.08	0.50	791	611	0.08	0.50
KIT-extrapolation_baseline	55,011	45,278	0.33	0.42	98,777	79,912	0.17	0.33	530	399	0.58	0.67	852	624	0.42	0.58
KIT-time_series_baseline	53,619	41,830	0.42	0.67	94,873	72,907	0.33	0.67	644	539	0.42	0.58	1,014	824	0.33	0.50
KITCOVIDhub-inverse_wis_ensemble	38,590	27,016	0.25	0.67	59,449	42,678	0.25	0.58	242	178	0.75	1.00	408	265	0.67	1.00
KITCOVIDhub-mean_ensemble	41,216	27,591	0.25	0.67	62,918	43,800	0.25	0.50	169	185	0.83	1.00	353	263	0.75	1.00
KITCOVIDhub-median_ensemble	41,979	28,852	0.17	0.58	61,177	43,683	0.08	0.58	196	173	0.75	1.00	354	256	0.75	1.00

*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 S3: Forecast evaluation for Germany and Poland, 1 and 2 weeks ahead (cumulative scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	1 wk ahead cumul case				2 wk ahead cumul case				1 wk ahead cumul death				2 wk ahead cumul death			
Model	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	9,252	5,415	0.25	1.00	29,303	18,479	0.25	0.75	300	204	0.50	0.92	739	496	0.42	0.92
epiforecasts-EpiNow2	14,097	11,023	0.67	0.75	42,806	32,126	0.58	0.75	588	457	0.67	0.83	1,130	804	0.58	0.83
FIAS_FZJ-Epi1Ger	10,859	6,690	0.50	0.92	35,351	21,697	0.17	0.83	578	458	0.00	0.25	976	739	0.25	0.50
Imperial-ensemble2									* 193	* 136	0.80	0.90				
itwm-dSEIR	7,517	5,379	0.42	0.67	26,189	19,526	0.33	0.67	726	548	0.25	0.50	1,149	972	0.33	0.50
ITWW-county_repro	15,223	12,418	0.08	0.25	46,473	37,889	0.08	0.33	564	527	0.00	0.00	813	742	0.08	0.17
Karlen-pypm	18,532	13,629	0.50	0.92	53,323	38,243	0.33	0.92	380	232	0.42	0.92	963	596	0.42	0.92
LANL-GrowthRate	6,889	9,267	1.00	1.00	21,087	22,068	0.83	1.00	466	284	0.25	1.00	866	532	0.42	1.00
LeipzigIMISE-SECIR	17,708	12,470	0.00	0.42	41,912	29,244	0.25	0.50	1,474	1,335	0.50	0.50	3,142	2,544	0.50	0.58
MIT_CovidAnalytics-DELPHI									851	504	0.33	0.92	1,533	969	0.08	0.75
SDSC_ISG-TrendModel	10,394								384							
USC-SIkJalpha	16,854	11,177	0.33	0.83	40,016	30,407	0.17	0.58	467	314	0.33	0.67	1,035	686	0.17	0.42
KIT-baseline	12,756	7,953	0.42	0.92	35,996	26,510	0.17	0.42	411	277	0.58	0.92	1,164	811	0.25	0.75
KIT-extrapolation_baseline	8,823	5,715	0.50	1.00	31,598	19,803	0.42	0.75	456	269	0.33	1.00	1,219	748	0.33	1.00
KIT-time_series_baseline	15,583	10,281	0.25	0.75	47,712	33,183	0.25	0.67	406	263	0.67	1.00	1,247	849	0.58	1.00
KITCOVIDhub-inverse_wis_ensemble	10,649	6,614	0.58	0.92	32,451	20,251	0.33	0.83	303	183	0.50	1.00	463	319	0.67	1.00
KITCOVIDhub-mean_ensemble	9,715	6,000	0.67	1.00	31,185	19,363	0.50	1.00	280	178	0.42	1.00	451	325	0.50	1.00
KITCOVIDhub-median_ensemble	8,118	5,344	0.50	1.00	27,596	18,028	0.42	0.92	283	161	0.50	1.00	544	342	0.50	1.00

Poland

	1 wk ahead cumul case				2 wk ahead cumul case				1 wk ahead cumul death				2 wk ahead cumul death			
Model	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	21,893	16,543	0.33	0.75	36,184	24,110	0.33	0.58	316	202	0.17	0.75	499	301	0.50	1.00
epiforecasts-EpiNow2	13,372	9,851	0.42	0.83	43,134	30,915	0.42	0.75	332	261	0.58	0.83	639	440	0.58	0.83
ICM-agentModel	*14,264	*12,390	0.64	1.00	*40,403	*30,044	0.64	1.00	*279	*279	0.91	1.00	* 624	* 735	1.00	1.00
Imperial-ensemble2									*188	*138	0.30	0.70				
ITWW-county_repro	20,054	17,364	0.17	0.25	56,144	48,119	0.17	0.42	589	551	0.00	0.00	1,374	1,276	0.00	0.00
LANL-GrowthRate	8,129	5,787	0.83	1.00	30,850	19,842	0.58	1.00	229	137	0.17	0.83	573	348	0.33	0.92
MIMUW-StochSEIR	5,933	4,132	0.33	0.83	23,176	19,204	0.08	0.25	251	238	0.17	0.25	524	497	0.00	0.00
MIT_CovidAnalytics-DELPHI									*320	*199	0.55	1.00	* 823	* 513	0.27	0.82
MOCOS-agent1	5,173	4,978	0.42	0.67	19,929	15,373	0.25	0.75	158	132	0.75	1.00	349	275	0.83	1.00
SDSC_ISG-TrendModel	12,372								201							
USC-SIkJalpha	10,405	6,919	0.33	0.83	42,376	32,924	0.08	0.42	206	133	0.33	0.92	441	279	0.17	0.75
KIT-baseline	16,407	9,736	0.42	0.83	44,384	32,767	0.17	0.42	258	167	0.42	0.92	664	457	0.17	0.50
KIT-extrapolation_baseline	9,448	5,992	0.50	0.92	38,410	27,616	0.25	0.75	269	190	0.58	0.83	672	466	0.50	0.83
KIT-time_series_baseline	10,784	7,787	0.75	0.83	41,180	29,143	0.42	0.75	300	232	0.67	0.67	735	593	0.50	0.67
KITCOVIDhub-inverse_wis_ensemble	8,743	5,630	0.42	0.92	28,915	18,765	0.42	0.83	131	105	0.83	1.00	320	231	0.67	1.00
KITCOVIDhub-mean_ensemble	8,448	5,465	0.67	0.92	28,569	18,187	0.42	0.83	125	108	0.83	1.00	274	235	1.00	1.00
KITCOVIDhub-median_ensemble	6,334	4,402	0.67	1.00	28,772	19,818	0.33	0.83	144	97	0.75	1.00	280	207	0.75	1.00

*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 S4: Forecast evaluation for Germany and Poland, 3 and 4 weeks ahead (cumulative scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	3 wk ahead cumul case				4 wk ahead cumul case				3 wk ahead cumul death				4 wk ahead cumul death			
Model	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	56,806	36,975	0.33	0.58	90,856	59,030	0.25	0.58	1,409	930	0.50	0.92	2,325	1,515	0.33	0.92
epiforecasts-EpiNow2	93,015	68,352	0.58	0.67	168,222	121,460	0.58	0.67	2,070	1,401	0.50	0.67	3,244	2,189	0.42	0.75
FIAS_FZJ-Epi1Ger	72,325	45,809	0.17	0.92	122,165	79,579	0.08	0.58	1,792	1,290	0.08	0.50	2,774	1,946	0.08	0.58
Imperial-ensemble2																
itwm-dSEIR	56,386	42,720	0.33	0.42	101,463	75,670	0.33	0.42	1,677	1,459	0.25	0.42	2,265	1,945	0.17	0.50
ITWW-county_repro	94,211	75,775	0.00	0.42	160,665	128,935	0.00	0.42	826	658	0.00	0.17	740	589	0.42	0.67
Karlen-pypm	107,997	76,940	0.33	0.83	195,806	138,671	0.17	0.75	1,896	1,172	0.08	0.83	3,205	2,033	0.00	0.83
LANL-GrowthRate	44,978	38,134	0.75	1.00	72,617	57,154	0.67	1.00	1,557	963	0.42	1.00	2,447	1,572	0.42	1.00
LeipzigIMISE-SECIR	81,273	56,348	0.33	0.58	146,917	102,340	0.33	0.75	4,929	3,727	0.33	0.67	6,882	5,009	0.08	0.67
MIT_CovidAnalytics-DELPHI									2,152	1,404	0.17	0.67	2,608	1,759	0.08	0.67
SDSC_ISG-TrendModel																
USC-SIkJalpha	72,202	62,065	0.08	0.17	115,293	105,336	0.08	0.25	1,796	1,368	0.08	0.25	2,744	2,223	0.00	0.08
KIT-baseline	70,752	56,271	0.08	0.25	114,908	96,625	0.00	0.17	2,352	1,760	0.17	0.42	3,867	3,094	0.08	0.25
KIT-extrapolation_baseline	67,400	42,285	0.25	0.75	118,558	76,508	0.17	0.75	2,572	1,608	0.17	0.75	4,493	2,909	0.17	0.75
KIT-time_series_baseline	94,122	67,918	0.17	0.58	153,758	115,049	0.33	0.58	2,553	1,830	0.50	0.92	4,372	3,271	0.50	0.83
KITCOVIDhub-inverse_wis_ensemble	66,373	41,830	0.33	0.83	111,839	72,513	0.25	0.67	893	571	0.58	1.00	1,535	958	0.50	1.00
KITCOVIDhub-mean_ensemble	63,145	40,346	0.50	0.83	108,994	70,062	0.42	0.83	972	599	0.42	1.00	1,684	997	0.25	1.00
KITCOVIDhub-median_ensemble	58,244	38,895	0.42	0.75	100,388	68,015	0.42	0.83	1,104	641	0.33	1.00	1,854	1,072	0.33	1.00

Poland

	3 wk ahead cumul case				4 wk ahead cumul case				3 wk ahead cumul death				4 wk ahead cumul death			
Model	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	71,493	46,562	0.17	0.67	124,129	86,643	0.08	0.50	801	495	0.50	1.00	1,188	756	0.50	0.92
epiforecasts-EpiNow2	97,781	68,679	0.33	0.75	180,004	125,246	0.25	0.58	1,189	794	0.58	0.92	1,985	1,354	0.42	0.92
ICM-agentModel	* 67,819	* 49,027	0.55	1.00	*101,414	* 73,220	0.55	0.91	*1,207	*1,262	0.91	1.00	*1,834	*1,789	1.00	1.00
Imperial-ensemble2																
ITWW-county_repro	118,193	101,211	0.17	0.33	224,258	192,827	0.08	0.33	2,410	2,187	0.00	0.00	3,854	3,406	0.00	0.08
LANL-GrowthRate	71,765	42,183	0.50	1.00	127,178	74,690	0.08	0.83	1,027	616	0.33	0.92	1,638	977	0.33	0.83
MIMUW-StochSEIR	53,052	46,717	0.08	0.25	106,465	96,945	0.00	0.00	832	787	0.00	0.17	1,186	1,072	0.00	0.17
MIT_CovidAnalytics-DELPHI									*1,540	*1,075	0.18	0.55	*2,558	*1,978	0.09	0.36
MOCOS-agent1	38,471	32,124	0.42	0.58	73,460	59,498	0.42	0.50	531	422	1.00	1.00	699	589	0.92	1.00
SDSC_ISG-TrendModel																
USC-SIkJalpha	84,104	75,355	0.08	0.25	145,603	136,915	0.08	0.08	625	428	0.33	0.58	992	701	0.17	0.50
KIT-baseline	85,544	69,035	0.08	0.33	137,648	116,676	0.00	0.25	1,286	965	0.00	0.33	2,063	1,658	0.00	0.33
KIT-extrapolation_baseline	93,229	72,727	0.33	0.50	184,714	148,616	0.33	0.33	1,203	859	0.42	0.67	2,021	1,465	0.42	0.67
KIT-time_series_baseline	95,539	70,731	0.50	0.75	186,952	142,431	0.42	0.67	1,336	1,132	0.50	0.50	2,342	1,966	0.42	0.58
KITCOVIDhub-inverse_wis_ensemble	64,825	42,213	0.33	0.83	116,123	80,227	0.25	0.67	563	381	0.75	1.00	979	629	0.67	1.00
KITCOVIDhub-mean_ensemble	66,331	42,082	0.33	0.83	118,378	80,761	0.33	0.75	416	386	1.00	1.00	750	617	0.75	1.00
KITCOVIDhub-median_ensemble	68,709	46,381	0.42	0.83	120,285	85,310	0.25	0.50	382	332	0.75	1.00	667	565	0.67	1.00

*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 S5: Forecast evaluation for Germany and Poland, 1 and 2 weeks ahead (incidence scale, based on JHU data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	1 wk ahead case (JHU)				2 wk ahead case (JHU)				1 wk ahead death (JHU)				2 wk ahead death (JHU)			
Model	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	13,319	8,491	2/12	9/12	22,391	15,479	3/12	8/12	336	226	4/12	11/12	485	323	4/12	11/12
epiforecasts-EpiNow2	14,041	10,182	4/12	10/12	31,439	22,548	6/12	8/12	284	196	9/12	11/12	601	424	7/12	9/12
FIAS_FZJ-Epi1Ger	14,650	9,740	2/12	8/12	27,022	17,565	1/12	9/12	389	299	2/12	5/12	636	468	2/12	8/12
IHME-CurveFit									461				661			
Imperial-ensemble2									*234	*155	7/10	9/10				
itwm-dSEIR	9,919	7,887	6/12	7/12	20,214	14,003	5/12	9/12	492	319	6/12	11/12	518	336	6/12	10/12
ITWW-county_repro	17,888	15,162	2/12	4/12	33,039	27,132	1/12	3/12	517	483	1/12	2/12	259	204	1/12	2/12
Karlen-pypm	22,754	16,774	4/12	10/12	34,263	25,765	4/12	10/12	393	253	6/12	11/12	656	414	2/12	10/12
LANL-GrowthRate	12,175	10,877	9/12	12/12	17,749	14,958	9/12	11/12	286	193	6/12	12/12	457	286	6/12	12/12
LeipzigIMISE-SECIR	14,283	11,458	2/12	5/12	28,741	20,567	2/12	6/12	406	292	7/12	12/12	912	656	4/12	8/12
MIT_CovidAnalytics-DELPHI	*15,629	*10,544	5/11	10/11	*23,970	*16,970	3/11	8/11	777	468	5/12	12/12	763	452	7/12	12/12
SDSC_ISG-TrendModel	9,981								419							
USC-SIkJalpha	18,702	12,899	2/12	9/12	26,340	18,768	1/12	6/12	393	260	7/12	10/12	573	343	3/12	9/12
KIT-baseline	16,125	10,728	4/12	10/12	26,070	18,921	1/12	6/12	423	291	7/12	11/12	814	544	2/12	8/12
KIT-extrapolation_baseline	13,419	8,526	4/12	10/12	24,159	15,031	3/12	11/12	450	282	5/12	12/12	826	506	4/12	10/12
KIT-time_series_baseline	20,128	13,215	1/12	8/12	34,678	23,736	2/12	7/12	467	295	7/12	12/12	844	610	6/12	11/12
KITCOVIDhub-inverse_wis_ensemble	12,749	8,145	5/12	11/12	24,149	14,899	4/12	11/12	185	148	8/12	12/12	331	213	10/12	12/12
KITCOVIDhub-mean_ensemble	12,822	8,185	6/12	12/12	24,164	14,858	4/12	11/12	202	150	7/12	12/12	360	224	7/12	12/12
KITCOVIDhub-median_ensemble	11,789	7,853	3/12	10/12	21,667	14,075	4/12	10/12	195	147	8/12	12/12	381	231	7/12	12/12

Poland

	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
Model	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	9,348	5,809	0.33	0.92	28,241	19,561	0.08	0.58	221	151	0.42	0.75	316	195	0.50	0.83
epiforecasts-EpiNow2	8,669	6,394	0.50	0.83	30,133	22,536	0.42	0.83	218	138	0.50	0.92	390	245	0.42	1.00
ICM-agentModel	*24,466	*16,579	0.27	0.91	*28,581	*19,023	0.73	1.00	*506	*296	0.73	1.00	*573	*504	0.82	1.00
IHME-CurveFit									352				472			
Imperial-ensemble2									*191	*143	0.30	0.60				
ITWW-county_repro	18,811	16,021	0.08	0.42	36,017	31,212	0.17	0.50	548	510	0.00	0.00	729	656	0.00	0.00
LANL-GrowthRate	10,738	6,547	0.67	1.00	26,194	16,557	0.42	0.92	249	146	0.08	0.83	378	236	0.25	0.83
MIMUW-StochSEIR	8,492	6,067	0.17	0.58	20,002	17,452	0.08	0.25	249	235	0.08	0.17	286	264	0.00	0.00
MIT_CovidAnalytics-DELPHI	*24,435	*14,346	0.20	0.90	*52,595	*35,605	0.10	0.40	*417	*260	0.45	0.91	*536	*314	0.27	1.00
MOCOS-agent1	8,086	6,272	0.50	0.58	17,320	13,263	0.25	0.50	159	134	0.67	1.00	159	141	0.83	1.00
SDSC_ISG-TrendModel	8,333								289							
USC-SIkJalpha	13,247	8,677	0.25	0.67	35,120	26,749	0.17	0.50	210	142	0.42	0.92	252	156	0.42	1.00
KIT-baseline	18,711	11,471	0.33	0.83	34,420	24,900	0.08	0.42	293	188	0.42	0.75	459	315	0.25	0.67
KIT-extrapolation_baseline	10,390	6,698	0.50	0.92	32,416	23,760	0.25	0.50	286	199	0.50	0.75	417	297	0.42	0.75
KIT-time_series_baseline	11,424	8,387	0.58	0.83	31,643	22,688	0.50	0.75	312	246	0.58	0.67	486	388	0.50	0.58
KITCOVIDhub-inverse_wis_ensemble	8,451	5,486	0.50	1.00	25,280	16,907	0.25	0.83	182	119	0.58	1.00	233	153	0.67	1.00
KITCOVIDhub-mean_ensemble	8,817	5,619	0.67	1.00	26,309	17,097	0.33	0.83	169	121	0.67	1.00	174	156	0.83	1.00
KITCOVIDhub-median_ensemble	8,977	5,411	0.50	0.92	25,951	17,939	0.17	0.75	190	114	0.58	1.00	215	142	0.75	1.00

*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 S6: Forecast evaluation for Germany and Poland, 3 and 4 weeks ahead (incidence scale, based on JHU data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	3 wk ahead case (JHU)				4 wk ahead case (JHU)				3 wk ahead death (JHU)				4 wk ahead death (JHU)			
Model	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	30,920	20,680	3/12	7/12	38,187	24,940	2/12	8/12	724	470	3/12	11/12	971	646	2/12	9/12
epiforecasts-EpiNow2	51,165	35,835	5/12	9/12	74,985	54,568	6/12	8/12	946	659	6/12	9/12	1,260	866	6/12	9/12
FIAS_FZJ-Epi1Ger	37,482	25,141	1/12	6/12	52,812	36,746	1/12	7/12	871	592	1/12	7/12	1,071	678	1/12	6/12
IHME-CurveFit									874				1,000			
Imperial-ensemble2																
itwm-dSEIR	31,973	19,974	4/12	10/12	47,231	30,301	5/12	10/12	532	335	5/12	12/12	539	337	5/12	12/12
ITWW-county_repro	48,002	38,724	1/12	5/12	64,365	53,500	3/12	5/12	330	246	2/12	4/12	781	625	1/12	1/12
Karlen-pypm	55,203	39,877	3/12	8/12	91,822	66,228	2/12	8/12	952	628	0/12	10/12	1,342	939	0/12	7/12
LANL-GrowthRate	23,138	17,800	9/12	11/12	32,959	22,202	6/12	12/12	722	457	5/12	11/12	1,076	675	6/12	11/12
LeipzigIMISE-SECIR	48,868	33,487	1/12	8/12	77,302	54,290	3/12	8/12	1,589	1,102	3/12	8/12	2,526	1,798	2/12	7/12
MIT_CovidAnalytics-DELPHI	*29,198	*22,885	5/11	7/11	*39,778	*31,367	5/11	7/11	718	465	9/12	12/12	680	488	9/12	12/12
SDSC_ISG-TrendModel																
USC-SIkJalpha	35,113	26,616	2/12	6/12	49,985	39,839	1/12	2/12	815	519	3/12	8/12	948	619	2/12	7/12
KIT-baseline	37,102	28,457	1/12	4/12	46,124	37,251	0/12	3/12	1,188	858	2/12	6/12	1,529	1,159	1/12	3/12
KIT-extrapolation_baseline	36,004	23,750	2/12	9/12	54,153	37,117	2/12	6/12	1,351	866	2/12	9/12	1,917	1,302	2/12	9/12
KIT-time_series_baseline	46,542	34,947	4/12	7/12	63,585	47,872	3/12	6/12	1,318	981	6/12	10/12	1,815	1,452	5/12	10/12
KITCOVIDhub-inverse_wis_ensemble	35,449	21,286	3/12	11/12	49,251	31,504	5/12	9/12	505	320	9/12	12/12	717	473	8/12	12/12
KITCOVIDhub-mean_ensemble	35,296	21,200	2/12	10/12	48,995	31,027	5/12	10/12	537	319	8/12	12/12	745	454	8/12	12/12
KITCOVIDhub-median_ensemble	31,963	19,924	3/12	10/12	46,236	28,865	3/12	9/12	553	338	4/12	12/12	727	475	5/12	12/12

Poland

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	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	49,190	36,126	0.00	0.42	66,413	49,981	0.08	0.42	381	248	0.58	0.92	538	316	0.42	0.92
epiforecasts-EpiNow2	57,000	40,754	0.25	0.58	84,239	61,040	0.25	0.50	590	408	0.58	0.92	858	598	0.42	0.92
ICM-agentModel	*36,110	*21,706	0.45	0.91	* 43,759	*28,736	0.27	0.64	* 580	*533	0.91	1.00	* 656	* 544	0.82	1.00
IHME-CurveFit									603				521			
Imperial-ensemble2																
ITWW-county_repro	64,635	56,004	0.08	0.50	111,127	96,150	0.00	0.25	1,020	886	0.00	0.08	1,444	1,219	0.00	0.17
LANL-GrowthRate	45,833	27,090	0.17	0.75	58,322	37,589	0.17	0.75	480	300	0.50	0.83	640	395	0.58	0.83
MIMUW-StochSEIR	37,055	33,387	0.08	0.25	61,665	56,148	0.08	0.08	347	314	0.17	0.17	422	350	0.17	0.17
MIT_CovidAnalytics-DELPHI	*77,200	*55,508	0.10	0.40	* 96,731	*73,059	0.10	0.30	* 770	*476	0.45	0.91	*1,062	* 698	0.27	0.64
MOCOS-agent1	25,951	21,656	0.25	0.42	43,679	32,808	0.17	0.42	210	164	0.92	1.00	220	203	0.92	1.00
SDSC_ISG-TrendModel																
USC-SIkJalpha	50,909	42,933	0.08	0.25	66,710	59,344	0.08	0.33	244	160	0.58	1.00	408	270	0.42	0.92
KIT-baseline	44,230	36,646	0.17	0.33	56,963	48,028	0.08	0.25	647	481	0.08	0.50	812	631	0.00	0.50
KIT-extrapolation_baseline	57,609	47,303	0.33	0.42	102,746	82,981	0.17	0.33	576	431	0.50	0.67	899	667	0.33	0.58
KIT-time_series_baseline	56,074	43,716	0.42	0.67	97,935	75,620	0.33	0.67	701	585	0.33	0.58	1,058	861	0.17	0.50
KITCOVIDhub-inverse_wis_ensemble	41,559	29,313	0.17	0.67	62,808	45,573	0.25	0.58	268	190	0.75	1.00	444	280	0.50	1.00
KITCOVIDhub-mean_ensemble	44,185	30,162	0.25	0.58	66,277	46,910	0.25	0.50	228	197	0.75	1.00	403	283	0.75	1.00
KITCOVIDhub-median_ensemble	44,949	31,671	0.17	0.50	64,535	46,726	0.08	0.58	229	187	0.75	1.00	399	278	0.58	1.00

*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 S7: Forecast evaluation at the regional level, Germany and Poland, 3 and 4 weeks ahead (incidence scale, based on RKI/MZ data). Results are averaged over the different regions (states in Germany, voivodeships in Poland). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	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-EpiNow2	2,697	1,803	0.33	0.73	3,879	2,630	0.29	0.70	98	66	0.44	0.80	136	95	0.42	0.78
FIAS_FZJ-Epi1Ger	2,840	1,929	0.19	0.58	4,192	2,862	0.18	0.52	75	54	0.16	0.45	96	65	0.13	0.51
IHME-CurveFit									64				68			
ITWW-county_repro	3,123	2,100	0.39	0.77	4,394	2,970	0.44	0.78	34	23	0.34	0.79	52	34	0.31	0.77
Karlen-pypm	3,634	2,261	0.43	0.84	5,740	3,578	0.30	0.75	64	36	0.41	0.92	84	49	0.33	0.85
LeipzigIMISE-SECIR			0.17	0.50			0.17	0.67			0.08	0.33			0.17	0.42
USC-SIkJalpha	2,190	1,584	0.23	0.65	2,830	2,094	0.19	0.54	53	36	0.35	0.82	65	45	0.30	0.81
KIT-baseline	2,194	1,636	0.17	0.50	2,774	2,168	0.11	0.40	79	54	0.25	0.69	96	69	0.20	0.57
KIT-extrapolation_baseline	2,496	1,629	0.21	0.69	3,392	2,315	0.16	0.60	114	76	0.36	0.78	157	106	0.35	0.75
KIT-time_series_baseline	2,806	1,989	0.20	0.61	3,571	2,748	0.21	0.55	132	90	0.38	0.78	178	130	0.32	0.73
KITCOVIDhub-inverse_wis_ensemble	2,622	1,675	0.38	0.82	3,679	2,375	0.35	0.79	55	35	0.41	0.93	77	51	0.39	0.89
KITCOVIDhub-mean_ensemble	2,503	1,582	0.37	0.86	3,496	2,254	0.37	0.80	61	39	0.44	0.93	86	58	0.44	0.89
KITCOVIDhub-median_ensemble	2,589	1,607	0.32	0.85	3,669	2,323	0.27	0.76	51	32	0.48	0.95	70	45	0.40	0.92

Poland

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	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-EpiNow2	3,849	2,745	0.36	0.69	5,889	4,209	0.27	0.61	77	63	0.45	0.83	149	194	0.41	0.83
ITWW-county_repro	4,319	3,222	0.37	0.73	7,133	5,353	0.29	0.66	71	53	0.20	0.60	98	70	0.26	0.69
USC-SIkJalpha	2,543	1,910	0.24	0.54	3,438	2,665	0.21	0.49	28	18	0.48	0.91	35	23	0.44	0.88
KIT-baseline	2,743	2,195	0.16	0.40	3,519	2,932	0.10	0.34	47	32	0.30	0.70	58	43	0.27	0.61
KIT-extrapolation_baseline	4,002	3,073	0.26	0.57	7,041	5,491	0.20	0.48	52	34	0.44	0.86	70	48	0.39	0.83
KIT-time_series_baseline	3,905	2,855	0.34	0.66	6,563	5,056	0.28	0.60	58	43	0.31	0.59	75	57	0.23	0.58
KITCOVIDhub-inverse_wis_ensemble	3,327	2,253	0.32	0.73	5,558	3,892	0.23	0.65	32	22	0.54	0.94	48	46	0.51	0.93
KITCOVIDhub-mean_ensemble	3,107	2,051	0.33	0.80	5,205	3,560	0.23	0.67	34	25	0.55	0.94	59	61	0.52	0.93
KITCOVIDhub-median_ensemble	3,301	2,095	0.35	0.76	5,412	3,441	0.24	0.62	30	20	0.54	0.92	40	28	0.51	0.89

*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 S8: Forecast evaluation for Germany and Poland, pooled across evaluation periods, 1 and 2 weeks ahead (incidence scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
Model	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	10,653	6,945	0.36	0.86	24,560	17,271	0.33	0.43	249	171	0.50	0.82	433	285	0.33	0.81
epiforecasts-EpiNow2	10,355	7,229	0.59	0.77	32,772	23,970	0.52	0.71	246	157	0.64	0.82	493	339	0.52	0.76
FIAS_FZJ-Epi1Ger	9,118	6,028	0.50	0.91	27,174	18,523	0.29	0.67	354	285	0.23	0.32	599	457	0.19	0.48
IHME-CurveFit																
Imperial-ensemble2									*224	*166	0.65	0.70				
itwm-dSEIR			0.42	0.75			0.42	0.75			0.58	0.83			0.50	0.83
ITWW-county_repro	23,951	19,913	0.05	0.23	45,782	37,544	0.00	0.19	476	449	0.05	0.09	393	342	0.10	0.24
Karlen-pypm			0.43	0.93			0.23	0.85			0.36	0.86			0.08	0.77
LANL-GrowthRate	*22,330	*15,271	0.74	1.00	*36,344	*23,394	0.61	1.00	*285	*187	0.47	1.00	*436	*281	0.50	0.94
LeipzigIMISE-SECIR	14,097	*13,366	0.24	0.65	37,135	*36,634	0.12	0.50	484	*318	0.41	0.76	828	*621	0.31	0.62
MIT_CovidAnalytics-DELPHI	*24,290	*17,004	0.37	0.74	*44,566	*33,677	0.33	0.56	*671	*436	0.30	0.70	*636	*397	0.37	0.89
SDSC_ISG-TrendModel	9,271								400							
UCLA-SuEIR																
USC-SIkJalpha	16,613		0.31	0.85	27,942		0.17	0.58	430		0.46	0.77	582		0.25	0.83
KIT-baseline	15,355	10,246	0.45	0.91	27,601	20,850	0.24	0.62	442	270	0.41	0.91	804	518	0.10	0.62
KIT-extrapolation_baseline	10,274	7,900	0.59	1.00	28,704	19,614	0.48	0.76	341	208	0.50	0.95	624	385	0.43	0.86
KIT-time_series_baseline	15,492	10,614	0.36	0.82	37,524	24,854	0.33	0.76	330	230	0.64	0.95	754	521	0.48	0.90
KITCOVIDhub-inverse_wis_ensemble	11,055	7,141	0.55	0.95	30,599	19,896	0.38	0.71	200	133	0.59	0.95	285	181	0.52	0.95
KITCOVIDhub-mean_ensemble	12,137	7,732	0.59	0.91	30,563	19,502	0.33	0.81	213	146	0.45	0.95	326	199	0.52	0.95
KITCOVIDhub-median_ensemble	9,249	6,221	0.59	0.95	27,149	17,998	0.38	0.81	217	143	0.59	0.91	358	221	0.48	0.90

Poland

	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
Model	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	10,292	6,836	0.59	0.86	30,493	20,653	0.19	0.62	243	155	0.36	0.95	423	264	0.33	0.86
epiforecasts-EpiNow2	9,327	6,422	0.55	0.82	33,681	23,474	0.33	0.76	276	184	0.45	0.86	670	465	0.43	0.76
ICM-agentModel	*19,736	*13,900	0.33	0.93	*32,194	*21,406	0.57	0.86	*599	*453	0.53	0.74	*1,101	*791	0.50	0.72
IHME-CurveFit																
Imperial-ensemble2									*293	*188	0.30	0.70				
ITWW-county_repro	19,188	16,147	0.18	0.32	35,214	29,629	0.19	0.48	560	521	0.00	0.05	748	669	0.00	0.10
LANL-GrowthRate	*11,012	* 7,151	0.68	1.00	*31,944	*19,233	0.44	0.94	*233	*151	0.32	0.89	* 366	*228	0.39	0.94
MIMUW-StochSEIR	*10,072	* 6,468	0.41	0.88	*23,492	*19,354	0.25	0.38	*422	*320	0.18	0.35	* 771	*557	0.06	0.12
MIT_CovidAnalytics-DELPHI	*27,212	*17,817	0.21	0.74	*54,488	*38,714	0.11	0.61	*447	*282	0.40	0.80	* 580	*354	0.26	0.89
MOCOS-agent1	8,855	6,863	0.32	0.59	22,131	17,207	0.14	0.52	174	139	0.82	1.00	296	202	0.81	0.95
SDSC_ISG-TrendModel	6,918								214							
USC-SIkJalpha	10,353		0.31	0.85	29,100		0.17	0.50	206		0.31	0.92	255		0.42	0.92
KIT-baseline	21,751	13,546	0.45	0.86	41,057	28,022	0.14	0.52	340	216	0.45	0.95	595	384	0.19	0.71
KIT-extrapolation_baseline	13,477	8,685	0.55	0.95	40,533	27,328	0.29	0.67	332	233	0.59	0.82	658	463	0.48	0.81
KIT-time_series_baseline	16,108	10,657	0.64	0.91	43,096	28,569	0.52	0.86	411	281	0.64	0.82	854	574	0.52	0.71
KITCOVIDhub-inverse_wis_ensemble	9,796	6,402	0.45	0.95	28,909	19,458	0.38	0.76	182	130	0.68	1.00	321	216	0.62	0.95
KITCOVIDhub-mean_ensemble	9,646	6,391	0.55	0.95	29,101	19,012	0.33	0.76	191	136	0.73	0.95	349	242	0.71	0.95
KITCOVIDhub-median_ensemble	10,342	6,430	0.55	0.95	30,228	19,832	0.29	0.76	186	124	0.64	1.00	312	202	0.52	1.00

*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 S9: Forecast evaluation for Germany and Poland, pooled across evaluation periods, 3 and 4 weeks ahead (incidence scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	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	36,332	25,966	0.25	0.55	48,205	32,980	0.05	0.63	673	449	0.25	0.70	936	677	0.21	0.63
epiforecasts-EpiNow2	70,236	51,689	0.40	0.75	134,886	98,969	0.37	0.63	826	574	0.45	0.75	1,336	906	0.42	0.74
FIAS_FZJ-Epi1Ger	53,281	37,314	0.15	0.65	88,229	64,186	0.16	0.63	877	648	0.05	0.50	1,046	737	0.16	0.53
IHME-CurveFit																
Imperial-ensemble2																
itwm-dSEIR			0.33	0.83			0.42	0.92			0.42	0.92			0.42	1.00
ITWW-county_repro	70,730	58,490	0.05	0.30	96,380	81,756	0.05	0.32	495	405	0.15	0.35	770	625	0.11	0.21
Karlen-pypm			0.33	0.75			0.17	0.67			0.00	0.75			0.00	0.67
LANL-GrowthRate	*45,698	*29,603	0.53	0.94	* 45,825	* 33,757	0.56	0.94	* 751	* 502	0.47	0.88	*1,044	* 711	0.50	0.88
LeipzigIMISE-SECIR	62,495	*72,724	0.07	0.53	102,043	*127,685	0.07	0.57	1,295	*1,108	0.20	0.60	2,003	*1,946	0.21	0.57
MIT_CovidAnalytics-DELPHI	*71,468	*58,057	0.28	0.56	*104,826	* 87,522	0.28	0.44	* 669	* 465	0.53	0.84	817	636	0.53	0.74
SDSC_ISG-TrendModel																
UCLA-SuEIR																
USC-SIkJalpha	41,039		0.08	0.42	53,915		0.17	0.25	808		0.25	0.67	1,007		0.25	0.67
KIT-baseline	38,805	30,289	0.25	0.45	48,062	37,794	0.16	0.37	1,188	842	0.10	0.40	1,536	1,186	0.05	0.21
KIT-extrapolation_baseline	54,987	36,555	0.30	0.85	93,240	64,624	0.21	0.58	1,099	711	0.30	0.80	1,758	1,169	0.16	0.79
KIT-time_series_baseline	65,037	45,753	0.30	0.70	98,051	78,394	0.37	0.68	1,216	905	0.40	0.75	1,762	1,545	0.37	0.79
KITCOVIDhub-inverse_wis_ensemble	56,618	38,957	0.25	0.75	94,005	65,809	0.32	0.74	473	301	0.65	0.95	687	458	0.47	0.89
KITCOVIDhub-mean_ensemble	53,170	35,696	0.20	0.70	81,675	57,091	0.32	0.74	506	310	0.45	0.95	699	434	0.42	0.95
KITCOVIDhub-median_ensemble	49,788	33,166	0.30	0.85	76,681	51,457	0.26	0.68	523	333	0.40	0.90	657	452	0.47	0.89

Poland

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	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	55,346	40,055	0.05	0.30	81,957	62,080	0.05	0.26	571	369	0.40	0.80	800	520	0.32	0.74
epiforecasts-EpiNow2	73,258	51,543	0.25	0.65	146,975	104,951	0.26	0.42	1,388	970	0.35	0.75	2,471	1,748	0.32	0.79
ICM-agentModel			0.46	0.77			0.33	0.58	*1,037	* 854	0.65	0.76	*1,020	* 846	0.62	0.81
IHME-CurveFit																
Imperial-ensemble2																
ITWW-county_repro	65,634	56,584	0.10	0.45	122,956	107,847	0.05	0.26	1,073	932	0.10	0.15	1,669	1,446	0.05	0.11
LANL-GrowthRate	*57,040	*31,430	0.12	0.88	* 72,916	* 45,076	0.12	0.75	* 530	* 329	0.41	0.94	* 774	* 479	0.56	0.88
MIMUW-StochSEIR	*42,959	*37,974	0.13	0.27	* 70,955	* 62,821	0.07	0.07	*1,403	*1,004	0.07	0.13	*2,131	*1,468	0.07	0.29
MIT_CovidAnalytics-DELPHI	*90,391	*67,331	0.12	0.41	*125,485	*100,228	0.19	0.31	* 875	* 589	0.28	0.83	*1,313	* 955	0.24	0.65
MOCOS-agent1	41,792	35,988	0.25	0.40	76,914	65,186	0.26	0.37	517	333	0.75	0.85	847	592	0.63	0.79
SDSC_ISG-TrendModel																
USC-SIkJalpha	42,643		0.08	0.33	51,952		0.08	0.33	344		0.50	1.00	530		0.42	0.92
KIT-baseline	55,156	42,440	0.10	0.40	67,121	52,517	0.16	0.37	854	615	0.15	0.50	1,068	828	0.11	0.47
KIT-extrapolation_baseline	83,345	62,296	0.35	0.55	165,052	127,853	0.21	0.37	1,116	845	0.60	0.70	2,046	1,585	0.47	0.58
KIT-time_series_baseline	81,657	61,117	0.50	0.80	144,083	125,616	0.47	0.79	1,280	918	0.45	0.70	1,777	1,322	0.37	0.68
KITCOVIDhub-inverse_wis_ensemble	54,734	39,215	0.20	0.60	96,497	74,679	0.26	0.53	560	366	0.55	0.90	1,070	701	0.42	0.89
KITCOVIDhub-mean_ensemble	56,238	39,027	0.25	0.65	97,337	73,756	0.26	0.53	567	395	0.55	0.90	1,045	705	0.47	0.89
KITCOVIDhub-median_ensemble	54,928	40,367	0.20	0.55	92,043	70,201	0.05	0.53	476	317	0.60	0.95	913	589	0.47	0.84

*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.