Collaborative Disease Forecasting in Real-Time: The Role of Nowcasting, Ensemble Models, and Evaluation Methods

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

(Dr.-Ing.)

von der KIT-Fakultät für
Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

M.Sc. Daniel Wolffram

Tag der mündlichen Prüfung: 08.04.2025

Referent: Jun.-Prof. Dr. Johannes Bracher
Korreferentin: Prof. Dr. Melanie Schienle

Karlsruhe, 2025

Acknowledgements

First of all, I would like to express my deepest gratitude to my primary supervisor, Jun.-Prof. Dr. Johannes Bracher, for his invaluable guidance, encouragement, and support throughout my PhD journey. I am also grateful to my supervisors, Prof. Dr. Melanie Schienle and Prof. Dr. Tilmann Gneiting, for their insightful feedback, invaluable contributions to my work, and for giving me the opportunity to conduct my research in a supportive work environment.

I am grateful to my colleagues at KIT, Prof. Dr. Fabian Krüger, Dr. Sebastian Lerch, Dr. Rebekka Buse, Dr. Konstantin Görgen, Dr. Lora Pavlova, Dr. Andreas Eberl, Lotta Rüter, Nils Koster, Jieyu Chen, Nina Horat, Barbora Sobolová, Friederike Becker, Lisa Leimenstoll, Tobias Biegert, Osama Warshagha, and Tobias Bodentien, for their collaboration, friendship, and valuable discussions. Despite extended home office periods, I greatly enjoyed your company and am thankful to know all of you!

Further, I would also like to thank my colleagues from HITS, Jun.-Prof. Dr. Timo Dimitriadis, Dr. Johannes Resin, Dr. Alexander Jordan, Dr. Benedikt Schulz, Dr. Jonas Brehmer, Dr. Eva-Maria Walz, Dr. Marc-Oliver Pohle, Dr. Ghulam Qadir, and Kristof Kraus, for fruitful collaborations and discussions.

I would like to express my sincere gratitude to Mirjam Wolf, Theda Schmidt, and Frauke Bley, whose organizational skills and assistance with administrative tasks made my research journey significantly smoother. I also want to thank the students I have supervised or worked with throughout the years: Jannik Deuschel, Christopher Bülte, Elisabeth Brockhaus, Jonas Littek, Davide Hailer, Jakob Ketterer, Yanting Liu, Sophia Seifert, Kai Reffert, Moritz Feik, Zhengting Wu, and Miro Ackermann.

I am grateful to have been part of HIDSS4Health, an interdisciplinary graduate school that broadened my understanding of data science in the context of health research. I am also thankful to all my international collaborators across various projects, without whose contributions and commitment this work would not have been possible.

On a more personal note, I am grateful for the support of my parents, Martina, Karl, and Jochen, as well as my brothers, Manuel and Freddy, throughout this journey. Their support, understanding, and belief in me, even from a distance, have meant a lot to me. Most importantly, I want to thank Rachel Ephivania, my wonderful wife and amazing mother to my sons, Jeremy (a.k.a. JJ) and Levin, who were my greatest source of joy during this journey. Lastly, I want to thank the final member of our family, our beloved dachshund Tony, who managed to keep me sane while simultaneously driving me insane. I can now confirm that training you is surely harder than finishing this entire thesis!

Abstract

Real-time infectious disease forecasting plays a crucial role in public health decision-making by improving situational awareness and informing resource allocation. However, accurate forecasting faces multiple challenges. Reporting delays distort real-time trends, requiring nowcasting methods to correct biases and provide a clearer picture of the current epidemic state. Meanwhile, short-term forecasting must account for rapidly evolving epidemic dynamics shaped by interventions, behavioral shifts, and emerging variants. In this complex and changing environment, probabilistic forecasts are essential for quantifying uncertainty and enabling informed decision-making. This thesis contributes to real-time epidemic surveillance by enhancing statistical methods for nowcasting and forecasting. Nowcasting techniques were systematically evaluated within the *COVID-19 Hospitalization Nowcast Hub*, demonstrating that nowcasts effectively reduce biases from reporting delays, particularly when combined into ensemble nowcasts. Probabilistic short-term forecasting was explored within the *German and Polish COVID-19 Forecast Hub*, similarly showing that ensemble forecasts provide better predictive accuracy and calibration than individual models. To further enhance predictive performance, an ensemble of data-driven models that integrate nowcasting and forecasting was developed for the *RESPINOW Hub*, a collaborative platform for real-time prediction of respiratory infections in Germany. Additionally, this thesis advances model evaluation by introducing novel methods and visual tools to enhance the assessment of probabilistic forecasts, particularly quantile forecasts. Coverage plots provide an intuitive representation of lower and upper coverage, making them especially valuable in low-count settings, while decompositions of divergence measures, such as Wasserstein and Cramér distances, allow for a clearer breakdown of forecast differences into interpretable shift and dispersion components. A key contribution of this work is the development and operation of the aforementioned hubs, which served as collaborative infrastructures for real-time epidemic prediction. These hubs facilitated automated data integration, acceptance and validation of submissions, ensemble computation, public visualization, and evaluation pipelines, ensuring that forecasting efforts remained transparent and accessible. By fostering open and reproducible science, this thesis ultimately strengthens the role of disease forecasting in epidemic surveillance and public health decision-making.

Contents

List of Figures

List of Tables

List of Abbreviations

AE	Absolute error
ARI	Acute respiratory infections
AVM	Area validation metric
CD	Cramér distance
CDF	Cumulative distribution function
CMIP	Coupled Model Intercomparison Project
CORP	Consistent, optimally binned, reproducible, PAV-based estimator
CRPS	Continuous ranked probability score
DSC	Discrimination component
ECDC	European Centre for Disease Prevention and Control
GEFCom	Global Energy Forecasting Competition
GPL	Generalized piecewise linear function
IDR	Isotonic distributional regression
JHU	Johns Hopkins University
MCB	Miscalibration component
MLP	Multi-layer perceptron
MZ	Polish Ministry of Health
NPI	Non-pharmaceutical intervention
OSF	Open Science Foundation
OxCGRT	Oxford Coronavirus Government Response Tracker
PAV	Pool-adjacent-violators algorithm
PI	Prediction interval
RKI	Robert Koch Institute
RSV	Respiratory syncytial virus
SARI	Severe acute respiratory infections
UNC	Uncertainty component
VOC	Variant of concern
WD	Wasserstein distance
WIS	Weighted interval score

1. Introduction

1.1. Motivation and relevance

The COVID-19 pandemic underscored the critical role of real-time epidemic modeling in guiding public health decision-making. Nowcasting provides a clearer picture of the current situation by correcting for reporting delays in epidemiological data, while short-term forecasting anticipates trends in the coming weeks. In contrast, scenario modeling explores "what-if" projections to inform long-term planning. Together, these approaches are essential for maintaining situational awareness, enabling efficient resource allocation during outbreaks, and supporting the planning of vaccination trials (Dean et al., 2020).

However, real-time epidemic forecasting comes with significant challenges, both methodologically and technically (Desai et al., 2019). Reporting delays obscure current trends, while forecasting models must continuously adapt to rapidly changing epidemic dynamics, influenced by non-pharmaceutical interventions (NPIs), behavioral shifts, and emerging variants. Scenario projections, on the other hand, rely on assumptions about these future developments, often rendering them unsuitable for rigorous evaluation against later observed data (Reich and Rivers, 2020).

Moreover, operationalizing real-time forecasts requires automated data pipelines that ensure continuous data retrieval, preprocessing, and validation. Forecasting systems must be able to handle evolving input data while applying real-time quality checks to detect missing values, inconsistencies, or formatting errors. In multi-team forecasting efforts, incoming forecast submissions must undergo automated validation checks before they can be integrated into a comprehensive forecasting framework. At the same time, ensemble forecasts must be updated automatically as new predictions arrive, while real-time visualization platforms ensure that the latest forecasts remain accessible. Addressing these technical challenges is critical to making real-time forecasts actionable and trustworthy.

A crucial aspect of short-term forecasting is the ability to quantify uncertainty, as outbreak dynamics often exhibit low predictability due to their complex and evolving nature (Held et al., 2017). Probabilistic forecasting, which provides entire distributions of possible outcomes rather than point predictions, is essential for expressing forecast confidence and enabling informed decision-making. Reliable uncertainty quantification allows public health officials to assess risk levels and plan accordingly.

At the same time, ensuring the reliability of these predictions requires robust evaluation tools. Statistical methods for assessing model performance, such as proper scoring rules and calibration diagnostics (Gneiting et al., 2007), are indispensable not only for improving predictive accuracy but also for fostering trust in results. Yet, evaluation strategies often lack accessibility or standardization, complicating efforts to compare and refine models across research teams (Nature Publishing Group, 2020).

The COVID-19 pandemic highlighted the importance of collaboration, transparency, and open science, leading to the emergence of forecasting hubs as powerful infrastructures where multiple independent teams contributed real-time predictions to a shared platform (Reich et al., 2022). These hubs fostered collaborative learning, reproducibility, and accessibility, ultimately enhancing the reliability of epidemic forecasts. However, designing and maintaining such infrastructures requires careful planning to ensure transparency, accessibility, and seamless integration of diverse modeling approaches and data streams.

Taken together, these challenges emphasize the need for methodological advancements in nowcasting and forecasting, improved evaluation frameworks, and collaborative infrastructures that promote openness and reproducibility in disease forecasting.

1.2. Contributions

This thesis presents a series of contributions to the field of real-time infectious disease forecasting, with a focus on both methodological advancements and practical applications. These contributions align with the three key objectives outlined in the previous section: enhancing nowcasting and forecasting, advancing model evaluation, and fostering open and collaborative science.

Enhancing nowcasting and forecasting

A key contribution of this thesis is the development and application of statistical techniques to improve nowcasting and forecasting in epidemic surveillance.

Nowcasting

Chapter 2 evaluates and compares various nowcasting methods applied to German 7-day hospitalization incidences, which became the guideline value for German pandemic policy during the COVID-19 pandemic. This study leveraged the newly implemented *Hospitalization Nowcast Hub*, a collaborative platform where multiple research teams applied nowcasting techniques to a shared dataset, enabling a comprehensive comparison of methods. The findings demonstrate that statistical nowcasting methods significantly reduce biases introduced by reporting delays. The study also highlights challenges in defining nowcasting targets and recommends ensemble nowcasts, which perform particularly well in predicting the most recent days – the most relevant target from a public health perspective.

Forecasting

This thesis advances forecasting methodologies by developing and evaluating multi-model and ensemble approaches. Chapters 3 and 4 focus on short-term forecasting of COVID-19 cases and deaths in Germany and Poland, conducted within the *German and Polish COVID-19 Forecast Hub*. This collaborative framework facilitated the comparison of probabilistic real-time forecasts produced by multiple independent teams.

A key contribution is the evaluation of ensemble forecasting techniques, including approaches that weigh models based on recent performance. Findings show that these weighted methods offer no clear advantage over simpler unweighted mean or median ensembles, which generally show good relative performance compared to individual models, especially in terms of coverage.

The studies further discuss challenges encountered during periods of rapid change, such as adapting to major trend shifts and the impact of emerging variants. The findings illustrate the importance of ensemble flexibility, as different models performed better in different epidemic phases. The work also highlights key differences in forecasting cases versus deaths, with death numbers proving more predictable due to their dependence on case and hospitalization trends. Additionally, these studies address challenges in forecast evaluation, particularly in handling truth data revisions and reporting inconsistencies, which complicate retrospective assessment and comparability over time.

Integrating nowcasting and forecasting

Chapter 7 presents the development of the *RESPINOW Hub*, a collaborative platform for both real-time nowcasting and forecasting of respiratory infections in Germany. Launched in fall 2024, the hub integrates diverse data streams into a unified system to support multiple pathogens, including seasonal influenza, respiratory syncytial virus (RSV), and pneumococcal disease, alongside other disease indicators.

As an initial step toward real-time deployment, a retrospective study was conducted to evaluate how integrating nowcasting into short-term forecasting improves predictions for hospitalizations due to severe acute respiratory infections (SARI). Nowcasting corrections were incorporated into multiple data-driven forecasting models, including a time series model, a gradient boosting method, and a neural network. These forecasts were then combined into an ensemble, which achieved the best overall performance. While the system produced well-calibrated predictions up to four weeks ahead, it struggled to anticipate an unprecedented double peak, which had not been observed in previous seasons. The study also provided insights into the effectiveness of machine learning approaches in epidemic forecasting, identifying both their strengths and limitations in this context.

Advancing model evaluation

Another key contribution of this thesis is in the area of model evaluation. It strengthens the methodological foundation for the rigorous evaluation of probabilistic predictions, which is essential for understanding their reliability and improving their utility in public health decision-making.

Evaluation methods and tools

Chapter 5 provides an overview of advanced evaluation tools for quantile forecasts, focusing on methods for assessing calibration and comparative evaluation. A central aspect of this work is the evaluation of calibration, which refers to the statistical consistency between predicted quantiles and observed outcomes. This thesis distinguishes between unconditional calibration, assessed through classical coverage criteria, and conditional calibration, which provides a more detailed view across different quantile levels and is visualized using quantile reliability diagrams. However, in discrete data settings, classical coverage criteria can be misleading, necessitating alternative approaches for assessing unconditional calibration. To address this, coverage plots are introduced as an intuitive tool for evaluating upper and lower coverage, offering clearer insights, particularly in low-count settings. Additionally, this work explores consistent scoring functions for the comparative evaluation of quantile forecasts, providing additional insights into forecast discrimination, miscalibration, and uncertainty.

Divergence measures

In Chapter 6, divergence measures such as Wasserstein and Cramér distances are used to quantify the differences between forecast distributions. This work contributes to the theoretical understanding of model performance by decomposing these distances into directed shift and dispersion components, which enhance interpretability and can be readily visualized for an intuitive assessment of forecast differences. These insights provide clearer guidance on how different models diverge, offering valuable information for model comparison and selection.

Fostering open and collaborative science

This thesis also underscores the importance of open science and collaboration in epidemic forecasting. A key contribution is the implementation and maintenance of multiple collaborative hubs, which served as essential infrastructures for real-time epidemic prediction.

Development and operation of hubs

Chapters 2, 3, 4, and 7 relied on dedicated collaborative hubs that were implemented as part of this work. The *COVID-19 Hospitalization Nowcast Hub* (Chapter 2) was established to compare nowcasting methods for German hospitalization incidences, operating with daily data updates and nowcast submissions. The *German and Polish COVID-19 Forecast Hub* (Chapters 3 and 4) facilitated real-time forecasting of weekly COVID-19 cases and deaths in Germany and Poland, accepting both national and subnational forecasts at the level of German federal states and Polish voivodeships, allowing for a more granular assessment of epidemic trends. The repository and code base developed for this hub served as a starting point for the *European COVID-19 Forecast Hub*, led by the European Centre for Disease Prevention and Control (ECDC), which expanded the scope to coordinate forecasts across multiple European countries. The *RESPINOW Hub* (Chapter 7) was developed to integrate nowcasting and forecasting for respiratory infections in Germany, supporting multiple pathogens, including seasonal influenza, RSV, and pneumococcal disease, as well as various disease indicators. This flexible and modular platform enables comprehensive real-time epidemic surveillance across different respiratory diseases.

All of these hubs required the setup and continuous operation of crucial infrastructure, including the management of data streams, preprocessing of input data, validation and acceptance of submissions, computation of ensembles, visualization of results via publicly accessible online dashboards, and final evaluation based on preregistered studies. These efforts established a robust and scalable framework for collaborative epidemic forecasting.

Reproducibility and collaboration

Beyond the technical implementation of these hubs, this work has contributed to advancing open science principles by ensuring reproducibility and transparency in forecasting. The preregistered studies, the open display of forecasts in real time, the publication of all results, and the structured evaluation pipelines across all hubs demonstrate best practices in reproducible research. The collaborative nature of these platforms allowed multiple research teams to compare, refine, and improve their models in real time, ultimately enhancing the reliability and credibility of epidemic forecasts. By fostering collaboration at scale, this thesis highlights the value of structured forecasting hubs in addressing complex public health challenges and ensuring that forecasting methods remain transparent and reproducible.

Summary

This thesis contributes to real-time disease forecasting by improving both nowcasting and forecasting techniques, advancing model evaluation through novel methods and tools, and establishing collaborative forecasting hubs as essential infrastructures for real-time prediction. These hubs facilitate data integration, multi-team forecast comparison, and reproducible evaluation pipelines, demonstrating the value of structured collaboration in improving predictive performance. By promoting open and transparent forecasting methodologies, this work strengthens the role of disease forecasting in public health decision-making and preparedness for future outbreaks.

1.3. List of publications

The chapters of this thesis correspond to the following publications. Due to the collaborative nature of these works, the specific contributions of the thesis author are listed below each paper for transparency.

Chapter 2

Wolffram, Daniel, Sam Abbott, Matthias An der Heiden, Sebastian Funk, Felix Günther, Davide Hailer, Stefan Heyder et al. "Collaborative nowcasting of COVID-19 hospitalization incidences in Germany", *PLOS Computational Biology* 19, no. 8 (2023): e1011394, https://doi.org/10.1371/journal.pcbi.1011394.

Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Chapter 3

Bracher, Johannes, **Daniel Wolffram**, Jannik Deuschel, Konstantin Görgen, Jakob L. Ketterer, Alexander Ullrich, Sam Abbott et al. "A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave", *Nature Communications* 12, no. 1 (2021): 5173, https://doi.org/10.1038/s41467-021-25207-0.

Implemented and maintained the forecast submission and processing system, assisted with evaluation analyses, and contributed to the interpretation of results and revision of the initial draft.

Chapter 4

Bracher, Johannes, **Daniel Wolffram**, Jannik Deuschel, Konstantin Görgen, Jakob L. Ketterer, Alexander Ullrich, Sam Abbott et al. "National and subnational short-term forecasting of COVID-19 in Germany and Poland during early 2021", *Communications Medicine* 2, no. 1 (2022): 136, https://doi.org/10.1038/s43856-022-00191-8.

Assisted in conceiving the study, implemented and maintained the forecast submission and processing system, contributed to the interpretation of results, assisted with evaluation analyses, and contributed to the revision of the initial draft.

Chapter 5

Gneiting, Tilmann, **Daniel Wolffram**, Johannes Resin, Kristof Kraus, Johannes Bracher, Timo Dimitriadis, Veit Hagenmeyer et al. "Model diagnostics and forecast evaluation for quantiles", *Annual Review of Statistics and Its Application* 10, no. 1 (2023): 597-621, https://doi.org/10.1146/annurev-statistics-032921-020240.

Conceptualization, Software, Visualization, Graphical tools, Methodology (lower and upper coverage, coverage plots), Data curation, Development of a data-driven ensemble model and real-time submissions to the US COVID-19 Forecast Hub, Writing - drafting and editing.

Chapter 6

Resin, Johannes, **Daniel Wolffram**, Johannes Bracher, and Timo Dimitriadis. "Shift-dispersion decompositions of Wasserstein and Cramér distances", submitted to *Statistical Science*, preprint arXiv:2408.09770 (2024), https://doi.org/10.48550/arXiv.2408.09770.

Conceptualization, Graphical tool for comparing model performance with interpretable shift and dispersion components, Application to climate data (Data curation, Formal analysis, Investigation, Software), Writing – review & editing.

Chapter 7

Wolffram, Daniel, Johannes Bracher, Melanie Schienle et al. "Integrating nowcasts into an ensemble of data-driven models for forecasting SARI hospitalizations in Germany", submitted to the *International Journal of Forecasting*, preprint medRxiv: 2025.02.21.25322655 (2025), https://doi.org/10.1101/2025.02.21.25322655.

Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Additional publications

The following publications, co-authored by the thesis author, are not directly part of this thesis but contribute to the broader research context.

Bracher, Johannes, **Daniel Wolffram**, Tilmann Gneiting, and Melanie Schienle. "Vorhersagen sind schwer, vor allem die Zukunft betreffend: Kurzzeitprognosen in der Pandemie", *Mitteilungen der Deutschen Mathematiker-Vereinigung* 29, no. 4 (2021): 186-190, https://doi.org/10.1515/dmvm-2021-0073.

Cramer, Estee Y., Yuxin Huang, Yijin Wang, Evan L. Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen **et al**. "The United States COVID-19 Forecast Hub Dataset", *Scientific Data* 9, no. 1 (2022): 462, https://doi.org/10.1038/s41597-022-01517-w.

Sherratt, Katharine, Hugo Gruson, Helen Johnson, Rene Niehus, Bastian Prasse, Frank Sandmann, Jannik Deuschel, **Daniel Wolffram** et al. "Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations", *eLife* 12 (2023): e81916, https://doi.org/10.7554/eLife.81916.

Brockhaus, Elisabeth K., **Daniel Wolffram**, Tanja Stadler, Michael Osthege, Tanmay Mitra, Jonas M. Littek, Ekaterina Krymova et al. "Why are different estimates of the effective reproductive number so different? A case study on COVID-19 in Germany", *PLOS Computational Biology* 19, no. 11 (2023): e1011653, https://doi.org/10.1371/journal.pcbi.1011653.

Amaral, André Victor Ribeiro, **Daniel Wolffram**, Paula Moraga, and Johannes Bracher. "Post-processing and weighted combination of infectious disease nowcasts", accepted at *PLOS Computational Biology* (2025), preprint medRxiv: 2024.08.28.24312701, https://doi.org/10.1101/2024.08.28.24312701.

2. Collaborative nowcasting of COVID-19 hospitalization incidences in Germany

2.1. Introduction

During infectious disease outbreaks, real-time surveillance data contributes to situational awareness and risk management, informing resource planning and control measures. However, the timely interpretation of epidemiological indicators is often hampered by the preliminary nature of real-time data. Due to reporting delays, the most recent data points are usually incomplete and subject to retrospective upward corrections. This bias can lead to incorrect conclusions about current trends. Statistical *nowcasting* methods aim to remedy this problem by predicting how strongly preliminary data points are still going to be corrected upwards, taking into account the associated uncertainty. Nowcasts thus help to uncover current trends which are not yet visible in reported numbers.

Problems of this type have been extensively researched across various disciplines; e.g., in econometrics, the gross domestic product and the inflation rate are routinely nowcasted (Giannone et al., 2008). Methods for preliminary count data as encountered in the present work originated in the actuarial sciences, where they were developed to handle insurance claims data (England and Verrall, 2002). In epidemiology, the problem of delayed reporting has been treated in diverse contexts, including the HIV pandemic (Cox and Medley, 1989), foodborne Escherichia coli outbreaks (Höhle and an der Heiden, 2014), the 2009 influenza pandemic (Donker et al., 2011) and mosquito-borne diseases like malaria (Menkir et al., 2021) and dengue (Bastos et al., 2019; McGough et al., 2020). During the COVID-19 pandemic, the problem has seen growing interest, and new approaches tailored to a variety of settings have been suggested (Günther et al., 2021; Seaman et al., 2022; Jersakova et al., 2022; Li and White, 2021; Hawryluk et al., 2021;

Greene et al., 2021). There is thus an ever-growing number of methods to statistically correct reporting delays. However, two important aspects are rarely addressed in the current literature. Firstly, few studies assess the performance of methods in real-time settings. The papers we are aware of – with Greene et al. (2021) as an exception – contain only retrospective case studies which risk smoothing over some of the difficulties occurring in real time (e.g., major data revisions, time pressure on analysts). Also, few studies include comparisons with existing methods. While occasionally one additional model is applied for comparison (McGough et al., 2020; Jersakova et al., 2022; Hawryluk et al., 2021), systematic comparative assessments are lacking. Our work fills this gap by examining multiple procedures in real time, thus providing a realistic picture of nowcast performance and the arising practical challenges. By bringing together several different models, our study is moreover the first able to assess the potential of combined ensemble nowcasts.

We evaluate the different nowcasting approaches in an application to German 7-day hospitalization incidences. These have played an important role in the management of the pandemic in Germany. Indeed, in November 2021, they were defined as the key indicator to determine levels of non-pharmaceutical interventions. Via a system of thresholds (German Federal Government, 2021), they played an important role in the management of the pandemic, in particular in the fall and winter of 2021. Nowcasting is of particular importance for this indicator due to the way it was defined. As will be detailed in Section 2.2.1, the official German hospitalization numbers published by Robert Koch Institute (RKI) are aggregated by the reporting date of the associated positive test rather than the date of hospital admission. The total time span between the case report and the hospitalization report (i.e. the "delay" that has to be predicted) thus consists of two parts: the time between the report of the positive test and hospital admission and the actual reporting delay between hospitalization and the reporting thereof. This definition led to some criticism in the public discourse but was defended as a necessary compromise between timeliness and data quality by RKI (Norddeutscher Rundfunk, 2021).

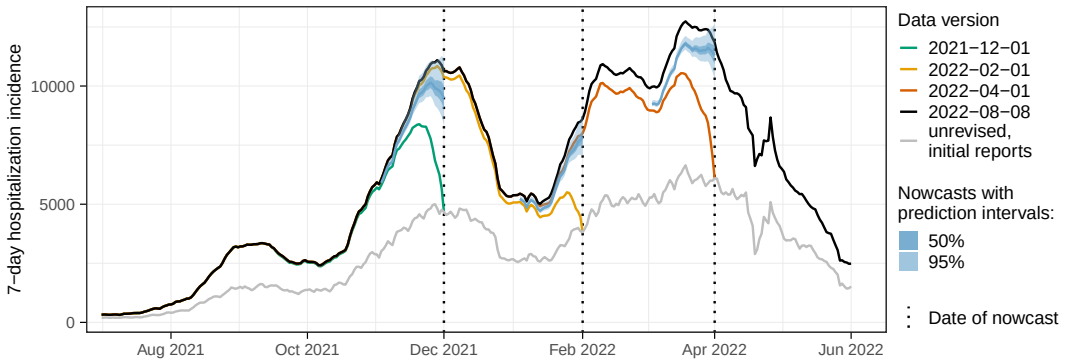


Figure 2.1.: Illustration of the nowcasting task. Data available in real time (colored lines) is incomplete, and especially for recent dates, the values are considerably lower than the final corrected values (black line). Nowcasts (blue-shaded areas) aim to predict in real time what the final data points will be. The light gray line shows the initially reported value as available on the respective date.

Figure 2.1 illustrates the nowcasting task in the context of the 7-day hospitalization incidence. It shows real-time nowcasts from 1 December 2021, 1 February 2022, and 1 April 2022. Comparison with a more stable data version from 8 August 2022 shows that in these instances, the nowcasts were able to correctly reveal the actual trends, which differed sharply from the apparent declines found in the data as available at the time of nowcasting.

The present work is based on a collaborative platform, the *COVID-19 Nowcast Hub*, which we launched soon after the hospitalization incidence became the guideline value for the German pandemic policy. It served to collect and combine real-time nowcasts from several models on a daily basis. The approach builds upon the *COVID-19 Forecast Hubs*, which during the pandemic were run in the US (Cramer et al., 2022b), Germany and Poland (Bracher et al., 2021b), and later the entire European Economic Area, Switzerland and the UK (Sherratt et al., 2023). These Hubs showed that combining different epidemiological models into an ensemble can produce more robust predictions, confirming results from forecasting challenges like *FluSight* on seasonal influenza (Reich et al., 2019a). We aimed for compatibility with the Forecast Hub ecosystem in many technical and methodological aspects, in particular by following the same submission format and evaluation criteria (Bracher et al., 2021a). This way we contribute to a growing evidence base on predictive epidemic modeling in real time.

The remainder of the manuscript is structured as follows. In Section 2.2, we introduce the 7-day hospitalization incidence as defined by RKI and outline the agreed-upon nowcast targets. We present the individual nowcasting methods and ensemble approaches, as well as the prespecified evaluation criteria. Section 2.3 presents the results of our formal performance evaluation, followed by qualitative observations on periods of unusual reporting patterns or the emergence of a new variant. We then assess the impact of model revisions as well as the sensitivity of the results to the exact definition of the nowcast target. Section 2.4 concludes with a discussion.

2.2. Methods

To facilitate a transparent assessment, we preregistered our evaluation study, specifying the criteria to assess the submitted nowcasts. The study protocol was deposited at the registry of the Open Science Foundation on 23 November 2021 (Bracher et al., 2021c). In some instances, we had to deviate from the protocol. These are detailed in the respective subsections and summarized in Table A.1 in the Appendix.

2.2.1. Definition of the COVID-19 7-day hospitalization incidence

Data on the German COVID-19 hospitalization incidence was published in a daily rhythm by Robert Koch Institute (Robert Koch Institute, 2023). By its official definition (German Federal Ministry of Health, 2021), it is given by the number of hospitalized COVID-19 cases among cases reported over a 7-day period relative to 100,000 inhabitants. As illustrated in Figure 2.2, hospitalizations are thus aggregated by the case reporting date, more precisely, when a case was digitally registered by a local health authority, rather than the date of hospital admission (though the two may coincide). We will refer to this case reporting date as the *reference date* in the following. We note that the hospitalization is not required to occur during the 7-day window mentioned previously, nor is COVID-19 required to be the main reason for hospitalization. When new hospitalizations are added to the record, they may thus change the value of the 7-day hospitalization incidence for past periods, depending on how much time has passed between the positive test, the time of hospitalization, and ultimately its reporting. Therefore, the initially reported

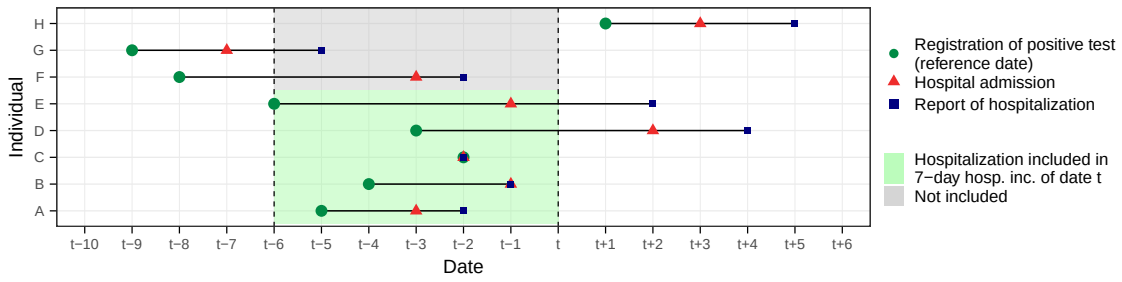


Figure 2.2.: Illustration of 7-day hospitalization incidences via individual-level timelines. The reference date by which hospitalizations are counted is the date when the positive test of an ultimately hospitalized person is reported (green dots). However, hospitalizations only become known after they take place (red triangles) and are reported (blue squares). Individuals A-E are included in the 7-day hospitalization incidence of date t because their reference date falls within a 7-day window from $t-6$ until t, even though some are reported as hospitalized later (individuals D and E). These hospitalizations only appear in the data with a delay and thus need to be predicted using a nowcasting method on day t. In principle, it is also possible that positive test, hospitalization and reporting all take place on the same day, as for individual C. In this case, there is no delay problem. We note that even though individuals F and G are hospitalized or reported within the period $t-6$ to t, they are not counted in the 7-day hospitalization incidence for day t because the positive test is reported before $t-6$. Individual H is not included because its reference date is after t.

value of the hospitalization incidence is merely an approximation and tends to be lower than the actual value.

To illustrate the extent of these revisions, Figure 2.3 shows the fraction of the 7-day hospitalization incidence that was reported 0 – 70 days after the respective reference date. Same-day values covered 50–60% of the ultimately reported hospitalizations, with a slight upward trend over the study period (left panel). Around 85% were reached after 14 days and even after 70 days, there were upward corrections of more than 3%. As illustrated in the second panel, same-day reporting completeness varied considerably across states. In Bremen (HB) it exceeded 75%, whereas it was below 50% in Saxony (SN) and Hamburg (HH). Reporting completeness was also variable across age groups and weekdays (third and fourth panels). A detailed display of temporal variations in initial reporting completeness across states can be found in Figure A.1 in the Appendix. It should be noted that initial reporting completeness can also depend on the overall

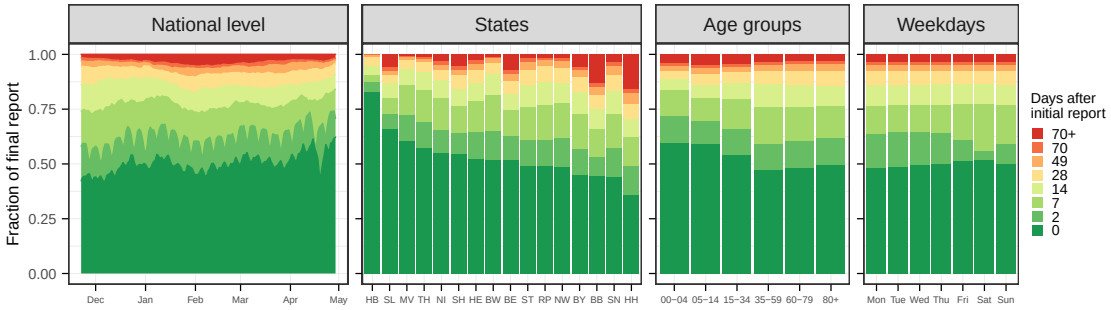


Figure 2.3.: Completeness of 7-day hospitalization incidences 0 to 70 days after the respective reference date. First panel: temporal development over the considered study period, aggregated over states and age groups. Second: by state, ordered by initial reporting completeness (see Figure A.1 in the Appendix for the definition of abbreviations). Third: by age group. Fourth: by weekday.

strain on the health system, and delays tend to be longer in times of high caseloads (Tolksdorf et al., 2022b).

As mentioned before, thresholds of 3, 6, and 9 per 100,000 population were introduced in the fall of 2021 and used to determine the necessary extent of non-pharmaceutical interventions (German Federal Government, 2021). These were applied to the initial value of the hospitalization incidence as reported on the respective day without any retrospective completion. This value is also referred to as the *frozen value*. For illustration, these frozen values were added to Figure 2.1 (light gray line). We note that due to the temporal and geographic differences shown in Figure 2.3, the same frozen value can translate to rather different final values of the hospitalization incidence.

2.2.2. Nowcast targets and study period

The goal of the collected nowcasts was to predict how much the preliminary values of the hospitalization incidence were still going to change. Specifically, on each day during the period from Monday 22 November 2021 to Friday 29 April 2022, a prediction needed to be issued for the final value the 7-day hospitalization incidence would take for that day and the previous 28 days. In the study protocol, we defined the final state to be predicted as the time series available on 8 August 2022. This date was chosen to be 100 days after the end of our study period. Originally, teams were asked to provide nowcasts

for all working days of the study period, excluding a Christmas break. However, as all teams fully automated their approaches, we were able to collect nowcasts on weekends and public holidays and include them in the study.

Teams were asked to issue nowcasts for the national level as well as for the 16 German states and seven different age groups (as available in public RKI data; 00-04, 05-14, 15-34, 35-59, 60-79 and 80+ years). No age-specific nowcasts at the state level were generated. To quantify prediction uncertainty, a probabilistic format was adopted, where teams had to submit seven quantiles (2.5%, 10%, 25%, 50%, 75%, 90%, 97.5%) of the predictive distribution in addition to the mean. Following the procedure in the various COVID-19 Forecast Hubs, our main analysis examined all outcomes on their original count scales, i.e. not standardized by population. This means that the relative size of states or age strata is reflected in the weight they receive in the overall evaluation (Bracher et al., 2021a).

In the study protocol, we also defined a retrospective study period reaching from 1 July 2021 to 19 November 2021. The motivation was to compare the retrospective performance on historical data available during model development to prospective performance under real-world conditions. However, due to time constraints, only two teams provided complete sets of retrospective nowcasts prior to the beginning of the prospective study. We therefore chose to omit this aspect. Instead, we added an evaluation of retrospective nowcasts from four revised models to the main study period from 22 November 2021 to 29 April 2022.

2.2.3. Overview of models

Nowcasts from eight independently run models were collected for the duration of our study. Six of them were contributed by groups of academics, one by the Robert Koch Institute (`RKI`) and one by the data science team of the newspaper *Süddeutsche Zeitung* (`SZ`). A short description of the different methods is provided in Table 2.1. Most approaches took preliminary hospitalization numbers as their only input, applying various techniques to model delay distributions and the underlying time series of hospitalizations. Only the `ILM` model took a different approach by including the number of confirmed cases as an explanatory variable. Approaches also differed in terms of the methods used for inference, uncertainty quantification, the flexibility and complexity of their delay distribution and

time series models, as well as the maximum delay considered (ranging from 35 to 84 days). Some models obtained nowcasts at a coarser spatial or age resolution by hierarchically aggregating nowcasts generated for finer strata. Models were typically not fitted to the entire available data set, but only a recent subset, the size of which again differed by team.

2.2.4. Ensemble approaches

On a daily basis, all submissions that were available at 2pm were combined to generate an ensemble nowcast, see Figure 2.4 for an illustration. We created the two following ensembles.

- For the `MeanEnsemble` each predictive quantile was obtained as the arithmetic mean of the respective quantiles of the member nowcasts. The ensemble mean was obtained as the mean of the submitted predictive means.
- For the `MedianEnsemble` the same procedure was applied using the median rather than the arithmetic mean for aggregation.

This direct aggregation at the level of quantiles rather than, e.g., probability density functions, is known as *Vincentization* (Genest, 1992). A discussion of its properties and differences to the aggregation of density functions can be found in Lichtendahl et al. (2013). As the expected number of contributed models was moderate, the `MeanEnsemble` was expected to be better-behaved than the `MedianEnsemble`, which can produce oddly shaped distributions in such settings (Bracher et al., 2021b). The `MeanEnsemble` was therefore prespecified as the primary ensemble approach (unlike in e.g. Cramer et al. (2022b) or Bracher et al. (2021b)).

We note that when using these aggregation approaches, quantile crossing can occur, meaning that the reported quantiles may not consistently increase with their nominal level (Koenker, 2005). To address this, one straightforward approach is to sort the quantiles in ascending order, which can improve the overall performance and coherence of the ensemble nowcasts (Chernozhukov et al., 2010). Regrettably, this consideration was overlooked in our real-time ensemble, leading to some instances of quantile crossing. These were primarily caused by missing entries for a specific quantile level in one of the

Table 2.1.: Description of nowcast models contributed to the collaborative project.

Abbreviation	Short description	Reference	Generation of prediction intervals	Data input	Weekday effects	Max. delay	Length of training data	Hierarchical aggregation
`Epiforecasts`	Bayesian model assuming the underlying curve of hospitalizations follows a random walk on the log scale. Reporting delays are assumed to follow a lognormal distribution with time-varying parameters. Report date effects are handled via a random effect for day of the week.	Abbott et al. (2021)	Bayesian posterior distribution	Hospitalizations	Yes	40 days	40 days	No
`ILM`	Hospitalization probabilities given a positive test are estimated separately per delay time and age group. These are used to predict yet unreported hospitalizations based on case incidences.	Heyder and Hotz (2023)	Based on past nowcast errors	Hospitalizations, cases	Indirectly via aggregation	84 days	91 days	Yes
`KIT`	A simple multiplication factor approach, with uncertainty intervals generated by comparing past point nowcasts to observations.	Section A.5 in the Appendix	Based on past nowcast errors	Hospitalizations	No	40 days	60 days	No
`LMU`	Nowcasts are based on a generalized additive model, with the delay distribution described by a sequential multinomial model.	Fritz et al. (2022); Schneble et al. (2021)	Parametric bootstrap approach using the covariance of the estimated model parameters and a Poisson observation model	Hospitalizations	Yes	40 days	56 days	Partially
`RIVM`	Counts per reference date and delay are modeled by a two-dimensional P-spline surface and covariates including weekday effects. This surface is extrapolated to fill in yet unknown values.	van de Kassteele et al. (2019)	Parametric bootstrap approach using the covariance of the estimated model parameters and a negative binomial observation model	Hospitalizations	Yes	42 days	84 days	No
`RKI`	The conditional reporting delay probabilities are modeled using a logistic regression model that incorporates weekday, federal state, and two age groups as covariates (≤ 60, > 60).	an der Heiden and Hamouda (2020); Lawless (1994)	Sampling from model-based distribution	Hospitalizations	Yes	40 days	68 days	Yes
`SU`	Similarly to `Epiforecasts`, the latent curve of daily hospitalizations is assumed to follow a random walk on the log scale. The delay distribution is modeled via a discrete-time hazard model with weekday effects for the reporting day. Entries of the reporting triangle are assumed to follow a negative binomial distribution.	Günther et al. (2021)	Bayesian posterior distribution	Hospitalizations	Yes	35 days	56 days	No
`SZ`	Nowcasts are based on the empirical distribution of the relative difference between initially reported and retrospectively completed values of the hospitalization incidence.	–	Empirical quantiles of past relative corrections	Hospitalizations	No	–	60 days	No

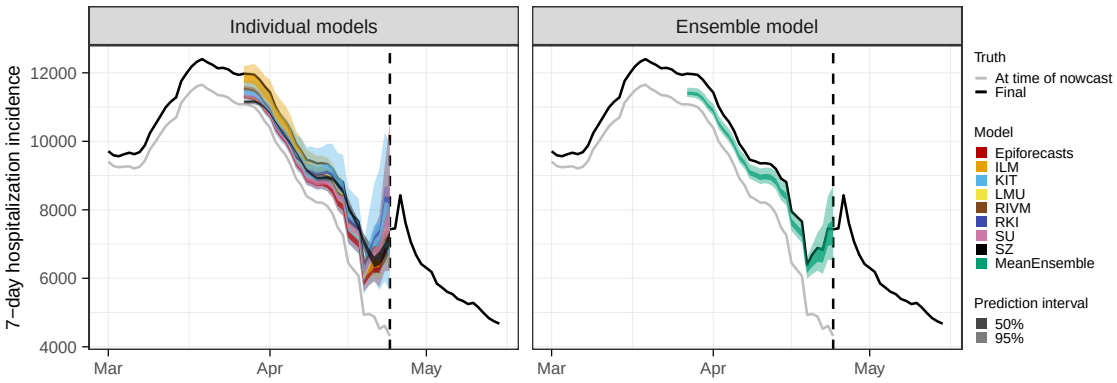


Figure 2.4.: Illustration of an ensemble approach. A set of individual nowcasts can be combined into an ensemble nowcast with different aggregation approaches. Here, the ensemble is computed as the quantile-wise mean of all nowcasts.

member models, as mentioned in Table A.4 in the Appendix. As this occurred only in a small fraction of instances, we consider the impact on overall results negligible.

2.2.5. Evaluation metrics

Proper scoring rules are an established tool to evaluate probabilistic forecasts (Gneiting and Raftery, 2007), or, in our setting, nowcasts. They are constructed such that they encourage honest forecasting, i.e., forecasters optimize their expected score by reporting their true beliefs about the future. Put differently, there is no way of "gaming" the system and obtaining improved scores by reporting modified versions of one's actual prediction. As in our setting nowcasts consist of three nested central prediction intervals, a natural choice is the interval score (Gneiting and Raftery, 2007). For an interval $[l, u]$ at the level $(1 - \alpha)$, $\alpha \in (0, 1)$, reaching from the $\frac{\alpha}{2}$- to the $(1 - \frac{\alpha}{2})$-quantile of the predictive distribution F, it is defined as

$$\text{IS}_\alpha(F, y) = (u - l) \; + \; \frac{2}{\alpha} \times (l - y) \times \mathbb{1}(y < l) \; + \; \frac{2}{\alpha} \times (y - u) \times \mathbb{1}(y > u), \qquad (2.1)$$

where $\mathbb{1}$ is the indicator function and y is the realized value. Here, the first term characterizes the spread of the predictive distribution, the second penalizes overprediction (observations fall below the prediction interval) and the third term penalizes underprediction. To assess all submitted quantiles of the predictive distribution jointly we use

the weighted interval score (WIS; Bracher et al. (2021a)), which is a weighted average of interval scores at different nominal levels and the absolute error. For N nested prediction intervals it is defined as

$$\text{WIS}(F, y) = \frac{1}{2N+1} \times \left(|y - m| \; + \; \sum_{k=1}^{N} \alpha_k \times \text{IS}_{\alpha_k}(F, y) \right), \tag{2.2}$$

where m is the predictive median and in our setting $N = 3$ and $\alpha_1 = 0.5$, $\alpha_2 = 0.2$, $\alpha_3 = 0.05$. We note that it is equivalent to the mean pinball loss across the respective quantile levels, which is often employed in quantile regression (Bracher et al., 2021a). The WIS approximates the widely used continuous ranked probability score (CRPS) and can be interpreted as a generalization of the absolute error to probabilistic predictions. It is negatively oriented, meaning that lower values are better. The decomposition of the interval score into spread, overprediction, and underprediction also translates to the WIS and can be used to enhance the interpretability of results.

To put results into perspective, we defined the simplistic baseline model `FrozenBaseline` which applies no correction and just issues the current data version as its deterministic nowcast (i.e., with all quantiles set to the same value). This allowed us to compute relative scores

$$\text{relative WIS of model } m = \frac{\text{mean WIS achieved by model } m}{\text{mean WIS achieved by baseline model}} \,,$$

characterizing the improvement over the uncorrected time series. Here, lower values are better, and values below 1 imply that the nowcasts reduce the error of the uncorrected time series. We note that while the study protocol specified that a baseline model was to be included, its definition was only agreed upon later. We note that the `KIT` model was originally conceived as a baseline model, but later considered too complex for this purpose; in the preregistration, it is therefore referred to as a "reference model".

To assess the central tendency of nowcasts we used the mean absolute error for predictive medians and the mean squared error for predictive means (i.e., for each functional we use the respective *consistent scoring function* (Ehm et al., 2016)). To evaluate calibration, i.e., the statistical consistency between nowcasts and observations, we consider the empirical

coverage of the 50% and 95% prediction intervals,

$$\text{coverage} = \frac{\#\text{ times nowcast intervals covered the final value}}{\#\text{ of nowcast intervals issued}} .$$

In case of missing submissions, i.e., if a team failed to provide a nowcast on a given day, nowcasts could be filled in retrospectively. To assess whether this had a substantial impact on the comparative evaluation, we applied a pairwise comparison scheme as described in Cramer et al. (2022b) to compare models using only the sets of nowcast tasks treated in real time by each model. Details can be found in Appendix A.4.

2.3. Results

2.3.1. Completeness of submissions

All participating teams produced nowcasts over the entire study period and only rarely failed to submit nowcasts in time (see Table A.2 in the Appendix). In most cases, missing nowcasts were filled in retrospectively. In very few cases (0.3% of all targets; see Table A.3 in the Appendix) it was not possible to obtain submissions from all teams; to handle these cases we chose to slightly deviate from the study protocol and omit the respective targets in our evaluation. The `ILM` model did not provide state-level nowcasts, while the `RKI` model did not include age-stratified results. Moreover, the `RKI` model only provided point nowcasts and two quantiles in real time (at levels 2.5% and 97.5%); the remaining quantiles were only provided in retrospect. We encountered some more minor difficulties, e.g., due to missing quantiles for certain targets; we summarize these and the chosen solutions in Table A.4 in the Appendix.

2.3.2. Visual inspection of nowcasts

For a first impression of nowcast performance, Figure 2.5 shows same-day nowcasts at the national level (i.e., at each date the respective nowcast with a horizon of 0 days is shown). Figure 2.6 shows the same for nowcasts 14 days back in time (i.e., for each day the nowcast issued 14 days later is shown). Displayed are the median predictions along with the central 50% and 95% prediction intervals. The light gray line shows the data as

available when the nowcast was issued (which in Figure 2.5 corresponds to the *frozen values*), and the red line shows the respective final value as available on 8 August 2022. In both figures, it can be seen that nowcasts from all models are generally close to the final values to be predicted. However, considerable variability in interval widths is apparent, ranging from rather wide (`KIT`) to very narrow intervals (`LMU`, `RKI`). Some models, in particular `KIT` and `SZ`, display pronounced weekday patterns in their same-day nowcasts, which to a lower degree also carry through to the ensemble nowcasts. For the nowcasts 14 days back in time we observe a slight downward bias in the central tendency, the only exception being the `ILM` model. As most of the concerned models moreover feature quite narrow prediction intervals, these often do not cover the final values.

2.3.3. Formal evaluation

To consolidate the qualitative findings from the previous section, we turn to a formal evaluation and consider evaluation scores and interval coverage rates. Figure 2.7 displays the mean and relative WIS values achieved by different models for the three considered aggregation levels (national level, states, and age groups). The left column shows mean scores (on the absolute and relative scale) across all strata and horizons, decomposed into contributions of spread, underprediction, and overprediction. The middle and right panels show the mean WIS and relative WIS by horizon, respectively (-28 to 0 days; see Section 2.2.5). At the national level and across age groups, the overall scores of the `ILM` model were considerably lower than the scores of all other models. The stratification by horizon indicates that it performed especially well for nowcasts seven or more days back. For the most recent days (-3 to 0 at the national level, -6 to 0 for age groups) the `MeanEnsemble` performed best. Across states, the `MeanEnsemble` outperformed the other models for horizons of -11 to 0 days. For horizons of -28 to -12 days, the `KIT` model achieved the best scores, which (by a narrow margin) led to the best overall result pooled across horizons. The relative scores indicate that, pooled over all horizons, most models were able to reduce the error of the uncorrected time series (`FrozenBaseline`) by roughly 80% (relative WIS of 0.2), while the `ILM` model achieved a reduction of about 90% (relative WIS 0.1). It is notable that `ILM` achieved almost constant improvements across horizons, while the improvements achieved by the other models were quite modest

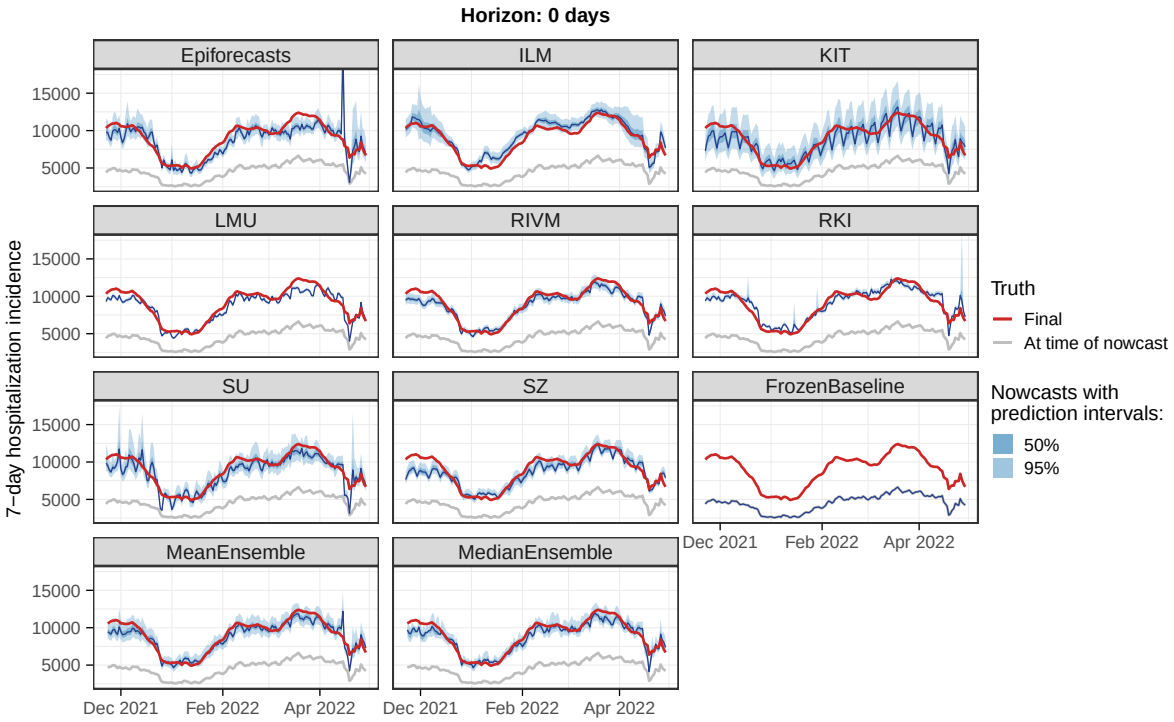


Figure 2.5.: Nowcasts with a horizon of 0 days back. Same-day nowcasts of the 7-day hospitalization incidence as issued on each day of the study period. Nowcasts are shown for the German national level.

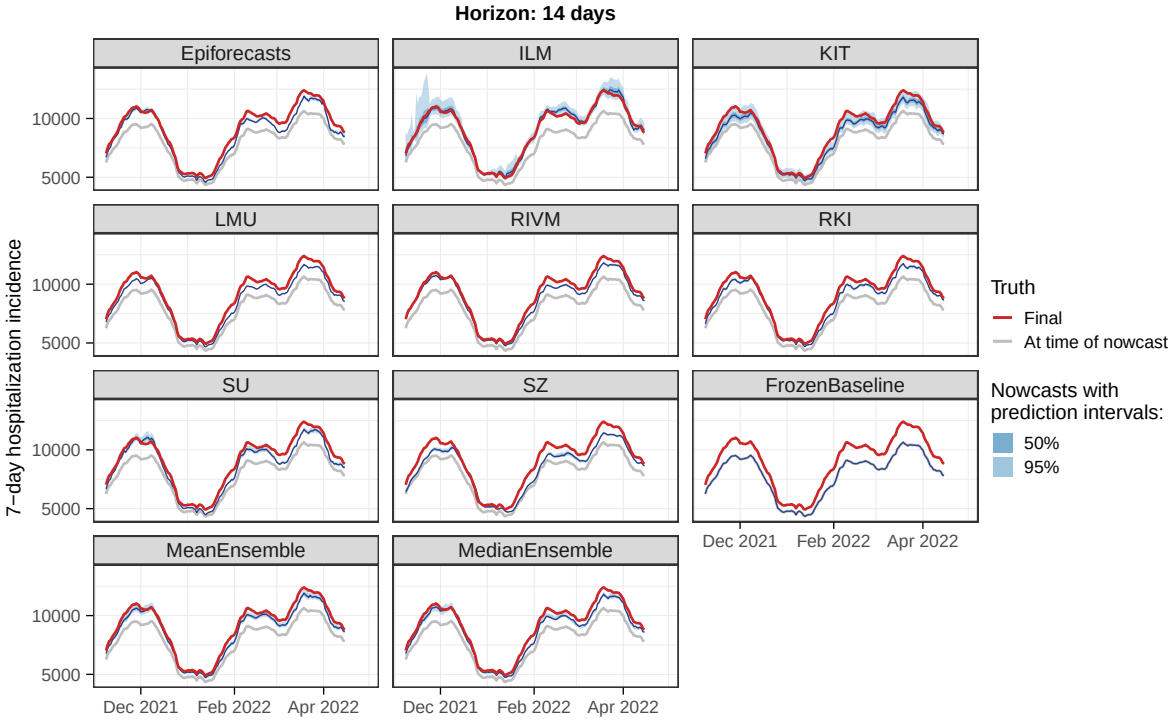


Figure 2.6.: Nowcasts with a horizon of 14 days back. Nowcasts of the 7-day hospitalization incidence as issued 14 days after the respective date. Nowcasts are shown for the German national level.

for horizons further into the past. To allow for a more detailed exploration of results we provide a display of the distribution of model ranks across individual nowcasting tasks (Figure A.7 in the Appendix) and of scores over time (Figure A.11 in the Appendix). Similarly to Cramer et al. (2022b), we find that the `MeanEnsemble` reliably achieved above-average performance across all locations and age groups (almost never ranking in the bottom). Additionally, stratified results by day of the week are shown in Figure A.8 in the Appendix. For `KIT` and `SZ`, which did not account for weekday effects, we indeed observe performance differences for different weekdays. Figure A.9 in the Appendix further shows nowcasts by `KIT` as issued on different weekdays, with a tendency for underprediction on Mondays and many cases of overprediction on Fridays or Saturdays. As seen in Figure A.10 in the Appendix, `ILM` on the other hand did not exhibit any such differences by weekday.

Figures A.4 and A.5 in the Appendix summarize results in terms of mean absolute errors of predictive medians and mean squared errors of predictive means in the same format as in Figure 2.7. The `ILM` model again performed favorably. Among the remaining models, `RIVM` shows good performance, in many cases outperforming the ensembles. The `KIT` model, on the other hand, which performed relatively well on WIS, achieved below-average results.

Empirical coverage rates of the 50% and 95% prediction intervals are displayed in Figure 2.8. Results are stratified by aggregation level (national, states, age groups) and nowcast horizon (-28 days to 0 days). The best calibration was achieved by the `ILM` model, with coverage rates close to the nominal levels at most horizons. Only for short horizons of -10 to 0 days coverage dropped moderately. In contrast, the `KIT` model achieved higher coverage rates for horizons between -4 and 0 days, which considerably dropped for nowcasts further into the past. All other models were overconfident and did not reach the respective nominal coverage levels. As for `KIT`, coverage was lower for nowcasts further back in time, for some models to a point where only a few observations were covered at -28 days.

To assess the impact of retrospective fill-in nowcasts for missing submissions, we recomputed relative WIS values using only real-time submissions and the pairwise comparison scheme from Cramer et al. (2022b). As can be seen from Table A.5 in the

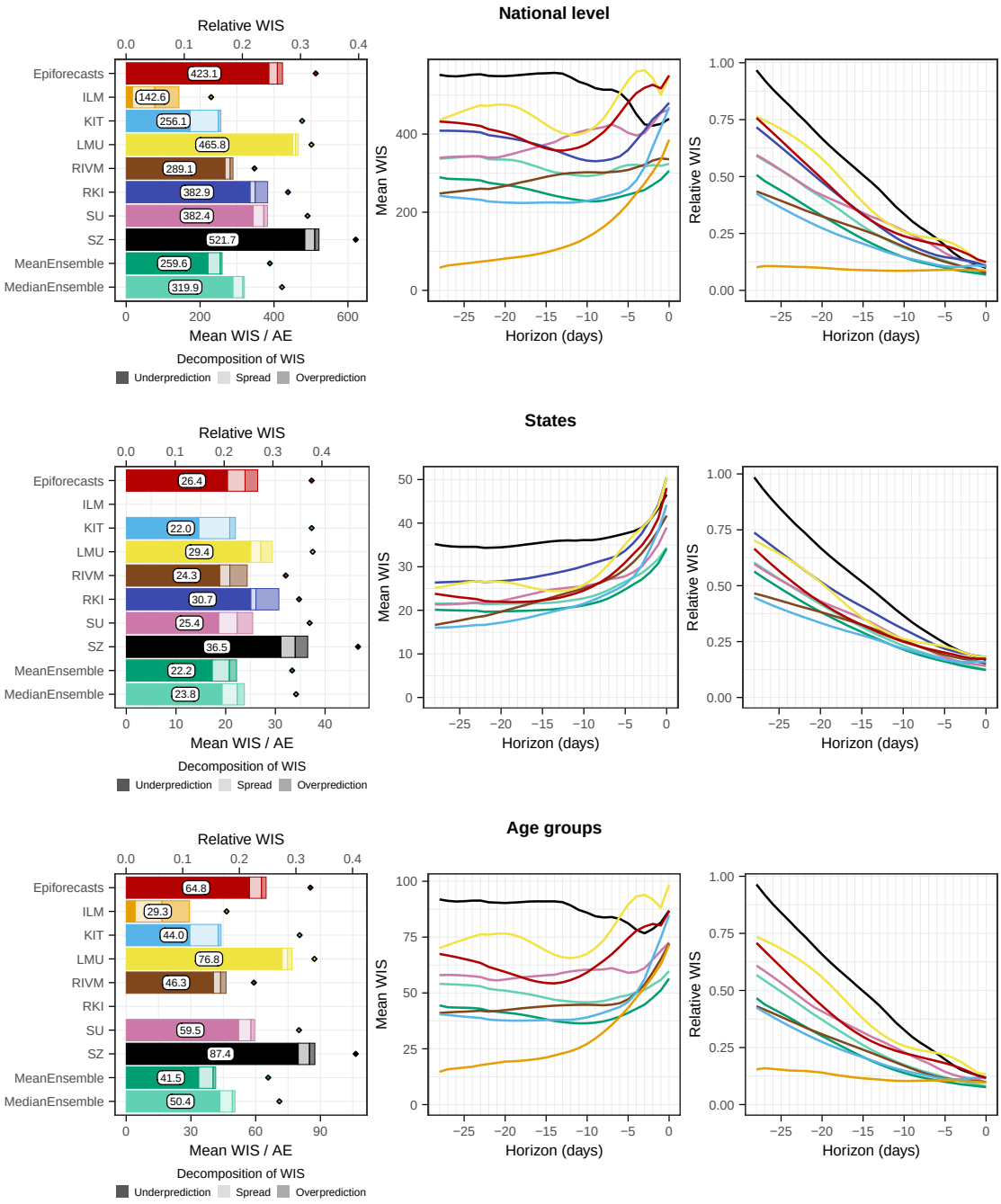


Figure 2.7.: Score-based performance. Shown is the mean WIS for the national level (top) and averaged across states (middle) and age groups (bottom). The first panel in each row displays the average across all horizons (on the absolute and relative scales). The decomposition into nowcast spread, underprediction, and overprediction (see Section 2.2.5) is represented by blocks of different color intensities. The absolute error is indicated by a diamond (⋄). The second and third panels in each row show the mean WIS and the relative WIS, respectively, stratified by horizon.

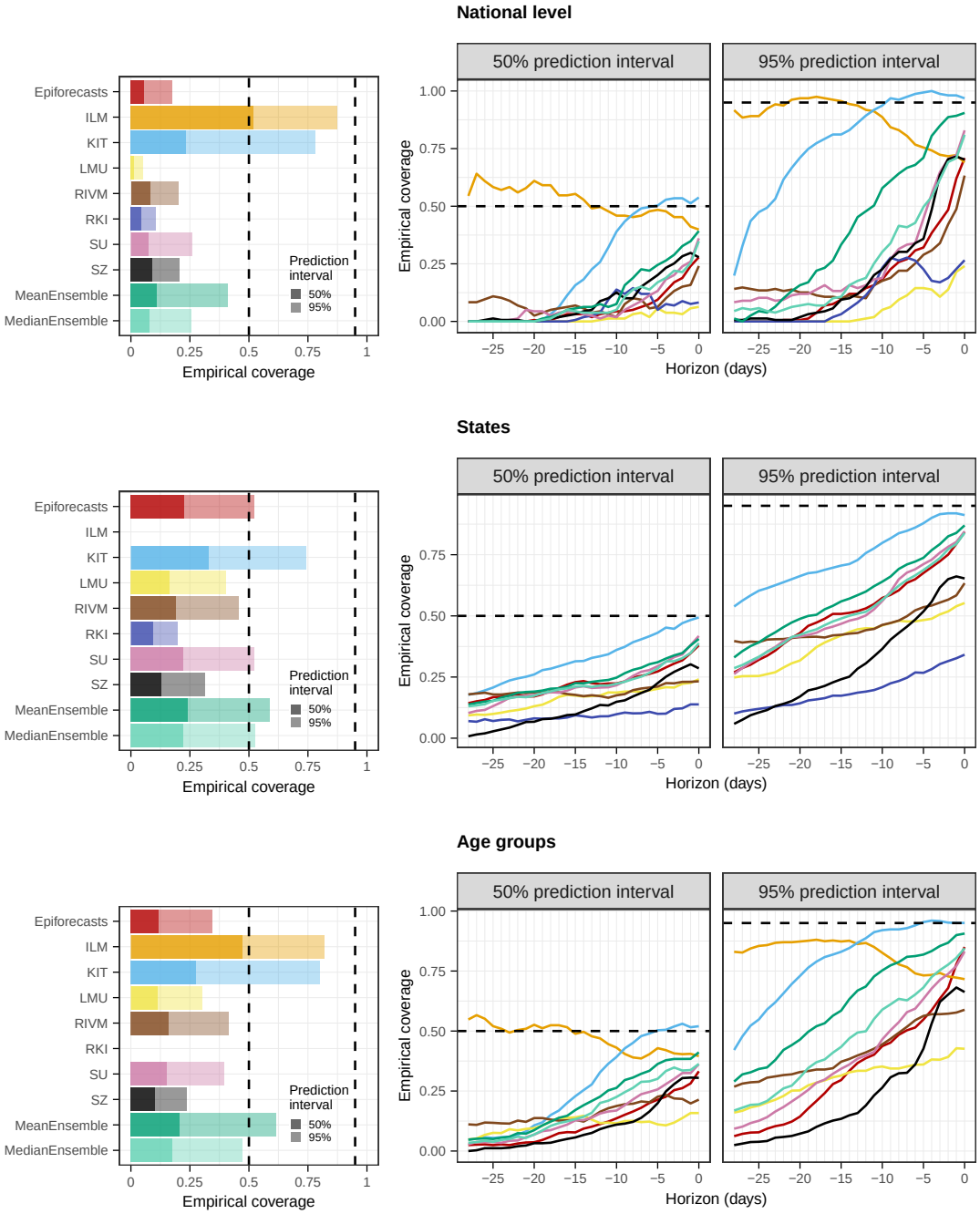


Figure 2.8.: Empirical coverage of the prediction intervals. Shown is the coverage for the national level (top), across states (middle), and across age groups (bottom). The first panel in each row displays the overall coverage of the 50% and 95% prediction intervals across all horizons. The second and third panels in each row show the empirical coverage of the 50% and 95% prediction intervals, respectively, stratified by horizon. The dashed lines indicate the desired nominal levels.

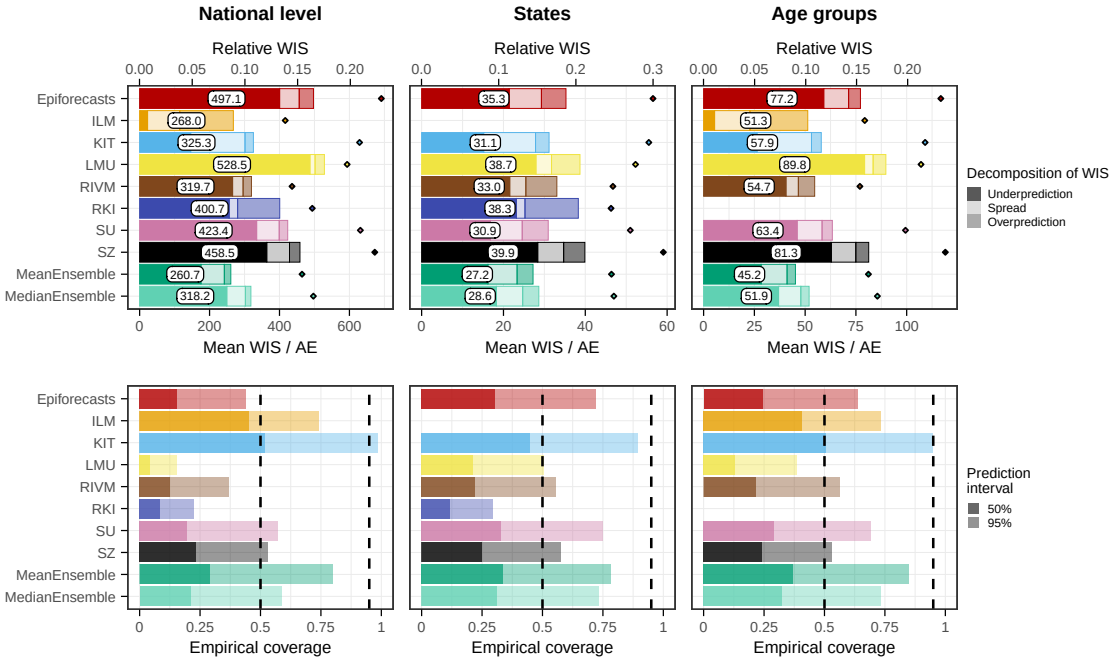


Figure 2.9.: Scores and coverage on short horizons. Shown are the mean WIS with absolute errors (top) and the empirical coverage (bottom) across horizons from 0 to -7 days.

Appendix the results barely change, indicating that fill-in nowcasts did not have a relevant impact on overall scores. Furthermore, as designated in the study protocol, we reran the evaluation using hospitalization incidences per 100,000 population rather than absolute hospitalization counts. The results are displayed in Figure A.12 in the Appendix and do not differ qualitatively from those in Figure 2.7.

As we consider the nowcasts for the most recent days the most relevant from a public health perspective, we conclude with an additional non-preregistered summary of scores across horizons -7 to 0 days. Figure 2.9 shows the average weighted interval scores and interval coverage rates. For this subset of nowcasting tasks, the `MeanEnsemble` outperformed the individual models in all three categories, closely followed by `ILM`. The `KIT` model reaches close to nominal coverage, while the other models are again overconfident.

2.3.4. Interpretation of evaluation results

As some of the presented results may seem contradictory at first sight, we provide some additional interpretations. Firstly, the opposing trends in absolute and relative WIS across horizons in Figure 2.7 can be interpreted as follows. All nowcasts – including the `FrozenBaseline` – get closer to the later observed final value as time passes and more complete data accumulates; thus, the absolute WIS decreases. However, most models seemed to have more difficulties predicting the small number of late additions than the bulk of early additions, leading to higher relative WIS. A possible explanation is that modelers needed to make a choice on which maximum delay to take into account. In light of Figure 2.3, the values of around 40 days as chosen by most teams may have been too low and led models to ignore a non-negligible fraction of hospitalizations still to be added. As can be seen from Figure 2.5, the resulting bias got largely absorbed in the overall uncertainty for same-day nowcasts. For the horizon of -14 days (Figure 2.6), on the other hand, it caused a visible shift between nowcasts and final values, which likewise led to insufficient coverage of prediction intervals.

The maximum delay chosen may also explain why the `ILM` model, which used a value of 80 rather than 40 days, was the best-performing individual method. However, the model also differed from the others in its general approach, using a regression on case incidences in addition to preliminary hospitalization numbers. We will attempt to shed more light on this aspect in Section 2.3.6. As a last relevant difference to most other models, the `ILM` approach based uncertainty intervals directly on the errors of past real-time nowcasts, an approach close to the idea of conformal prediction (Shafer and Vovk, 2008). A similar approach was also taken by the `KIT` model (see Section A.5 in the Appendix). The fact that these two models achieved the best calibration indicates that this approach may quantify nowcast uncertainty more realistically than standard model-based uncertainty intervals.

The decomposition of the WIS into components for spread, overprediction, and underprediction (Bracher et al., 2021a), displayed in the left column of Figure 2.7, is informative on the challenges the different approaches faced. Penalties for underprediction make up a very large part of the overall scores for all models except for `ILM`. This confirms the observation of a downward bias from Figure 2.6.

The best-performing individual models `ILM` and `KIT` issued predictive distributions with higher variability than the other models, indicated by the larger spread component. As can be seen from the diamond symbols in Figure 2.7 and in more detail from Figure A.4 in the Appendix, the `KIT` model did not issue particularly accurate point predictions (predictive medians). A likely reason is the lack of weekday effects, see Figures A.8 and A.9 in the Appendix. Its lower WIS values were primarily a result of better uncertainty quantification.

2.3.5. Impact of unusual reporting patterns and changes in virus properties

The nowcasting models in our study assumed either that the probability of hospitalization given a positive test remains roughly constant (the `ILM` model) or that the delay distribution in hospitalizations does so (all other models). In Figure 2.10, we therefore show four examples where these assumptions were violated. In mid-November 2021, hospitals in Saxony were overwhelmed (Berliner Morgenpost, 2021), leading to disruptions in the reporting system. As a consequence, initial reporting completeness dropped rather suddenly. This led the majority of models to underpredict, leading to an ensemble nowcast that was too low, as illustrated in Figure 2.10A. We note that in this instance we were aware that nowcasts for Saxony were likely unreliable and issued a warning on our website. Figure 2.10B shows an unusual reporting pattern from the state of Bremen from early 2022. Here, a relevant number of reported hospitalizations got removed from the record on 12 and 13 January, presumably due to faulty initial reporting. Nowcasts issued up to 11 January were thus considerably above the final data version from 8 August. Figure 2.10C and 2.10D show issues arising after the Easter weekend of 2022, when initial reporting was considerably lower than usual. As can be seen in Figure 2.10C, this led to too low ensemble nowcasts on Tuesday, 19 April. Also, over the following days, it seems to have caused issues in the fitting of certain models. As an example, Figure 2.10D shows the `Epiforecasts` output for Lower Saxony on 20 April. It features an excessively wide prediction interval, likely as a reaction to rapidly shifting delay distributions in the previous days. The `MeanEnsemble`, shown in green, was strongly affected by this

unusual behavior of a member nowcast, while the more robust `MedianEnsemble` remained unaffected.

A last noteworthy particularity is the behavior of the `ILM` model in January 2022, following the transition from the Delta to the Omicron variant. The Omicron variant is known to have lower clinical severity than the Delta variant (Wolter et al., 2022), meaning that during the transition the ratio of hospitalizations and confirmed cases gradually declined. For the `ILM` model, which assumes this ratio to remain constant, this led to an upward bias in nowcasts, which can be discerned from Figure 2.5 as an upward bump not present in the other models.

2.3.6. Retrospective variations of models

We next aim to shed some more light on how different modeling choices impact performance and how learnings from our study period facilitated the improvement of methods. To this end, we assess the performance of four variations of previously discussed models, which were applied retrospectively:

- The `LMU` team implemented a new approach to generate uncertainty intervals, which like the `ILM` and `KIT` methods is based on past nowcast errors.
- The `RKI` team obtained nowcasts by aggregating over finer strata and originally assumed independence across strata to generate prediction intervals at the aggregate level. This was changed to an assumption of strong correlations across strata, leading to wider prediction intervals.
- The `KIT` team reran its model with an increased maximum delay of 80 days, in contrast to the 40 days used in real time. This also required an increased length of training data, which was set to 100 days.
- Conversely, the `ILM` model was rerun with a maximum delay of 42 rather than 84 days, which is comparable to the maximum delays used by the remaining models. This was not meant as an improvement but as an adjustment to assess the impact of longer/shorter maximum delays.

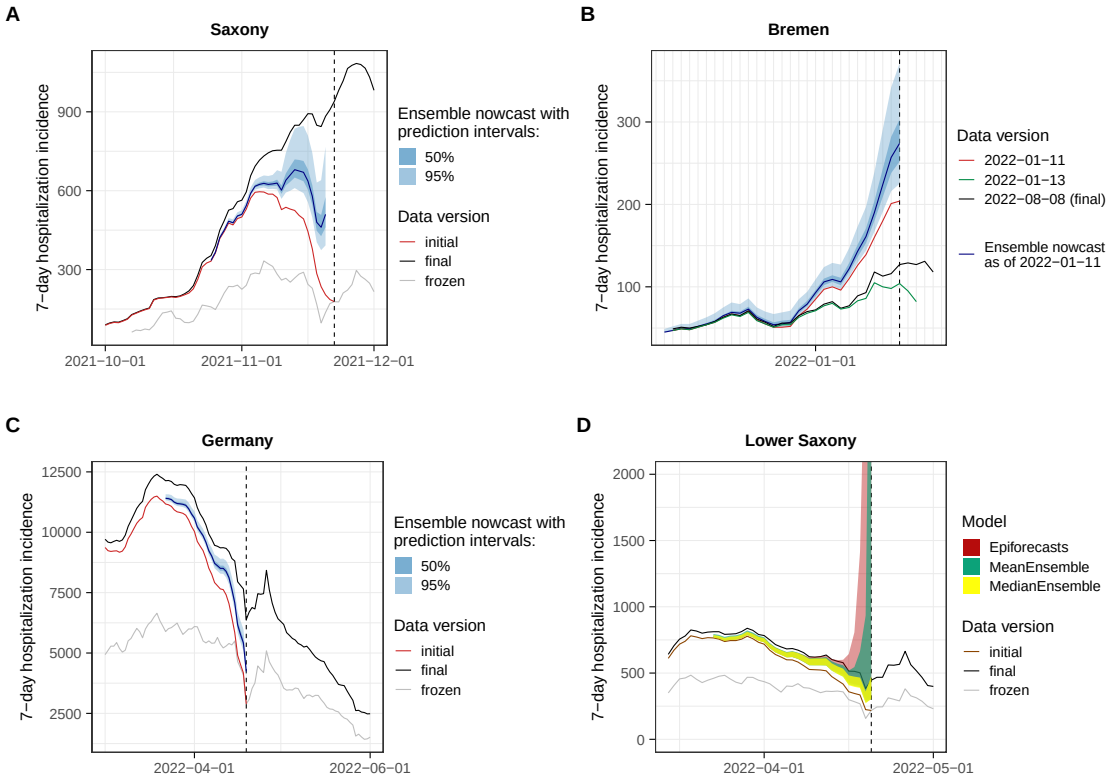


Figure 2.10.: Examples of time points when delay distributions were subject to sudden changes. **(A)** Saxony, nowcast made on 22 November 2021: overwhelmed hospitals lead to severe underreporting and thus too low nowcasts. **(B)** Bremen, nowcast made on 11 January 2022: some incorrect entries got removed from the records, resulting in a downward correction and thus too high nowcasts. **(C)** Germany, nowcast made on 19 April 2022: following the Easter weekend with lower than usual initial reporting coverage, nowcasts were considerably too low. **(D)** Lower Saxony, nowcasts made on 20 April: after the Easter weekend, `Epiforecasts` issued very wide nowcast intervals, presumably due to numerical problems. The dashed lines indicate the time when the nowcasts were made. (Incidentally, in example (A), the horizons of 0 days and 1 day back were missing, see Table A.3 in the Appendix.)

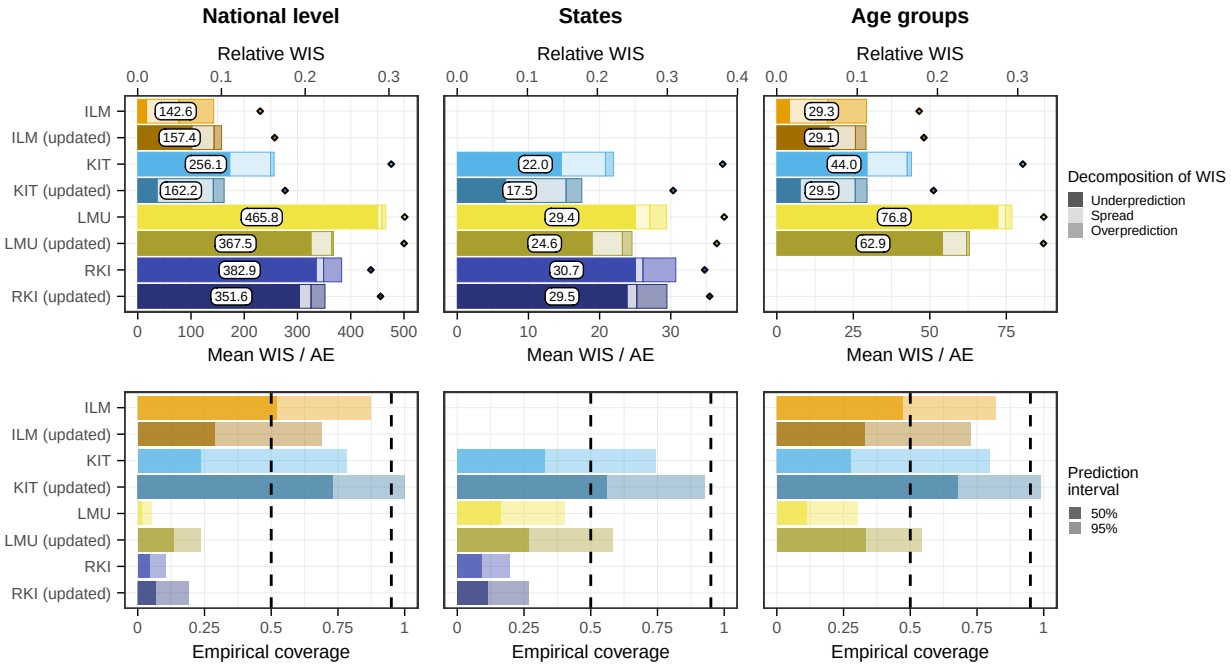


Figure 2.11.: Evaluation of retrospective model variations. Comparison of variations of the `ILM`, `KIT`, `LMU`, and `RKI` models and the same models as submitted in real time. Shown are the mean WIS with absolute errors (top) and the empirical coverage (bottom). Results are comparable to those from Figures 2.7 and 2.8.

The results are shown in Figure 2.11. They indicate improvements across all aggregation levels and horizons for the revised `LMU`, `RKI`, and `KIT` models. In particular across age groups, the `KIT` model now came close to the performance the `ILM` model achieved in real time. The coverage proportion of prediction intervals was increased for all three models, with the updated `KIT` model even leaning towards over-coverage (too wide intervals). The `LMU` and `RKI` models, on the other hand, remained overconfident. Decreasing the maximum delay in the `ILM` model slightly reduced the overall performance on the national level, while average scores across age groups remained almost unchanged. The score decomposition shows that the adjusted model tended to underpredict (similarly to the other models), while the original model tended to overpredict. Possible explanations will be discussed in Section 2.4.

2.3.7. Sensitivity of results to definition of final data

In our study protocol, we specified that the final state of the time series to be predicted was the version available on 8 August 2022, i.e., 100 days after the end of the study period. However, as we became aware of the fact that data revisions could occur with considerably longer delays than initially expected, we performed a sensitivity analysis to assess the impact of this choice. Figure 2.12 shows how the average WIS aggregated over horizons and different levels of stratification (i.e., the results shown in the left column of Figure 2.7) change when using a different data version as the final one. It can be seen that the average scores of all models, except for `ILM`, increase in parallel as newer data versions are used. The increase for `KIT` is slightly more gradual. This is because these models tend to underpredict, and as time passes and more additions are made to the data, this problem is exacerbated. For the `ILM` model, which tends to overpredict, average scores initially decline and then plateau, leading to an even more pronounced lead relative to the other models. As can be seen from Figure A.6 in the Appendix, using a later data version for evaluation, `ILM` ultimately also surpasses the ensemble when restricting results to horizons 0 to 7 days back.

The original target allowed revisions until the final date of 8 August 2022, meaning that for different reference dates, adjustments could be made over time periods of different lengths (e.g., the hospitalization incidence for the first reference date in our study period could be completed over a longer time than that of the last reference date). Denoting the hospitalizations for reference date t reported with a delay of d days by $x_{t,d}$, the nowcasting target y_t can formally be written as

$$y_t = \sum_{i=0}^{6} \sum_{d=0}^{t_{\max}-(t-i)} x_{t-i,d} \quad \text{with } t_{\max} = \text{2022-08-08},$$

where the first sum represents a 7-day window ending in date t and the second sum accumulates the hospitalizations with reference date $t - i$ reported until $t_{\max}$.

An alternative to choosing one specific data version as the nowcast target is a "rolling" approach that considers delayed reports for each reference date t up to a specified

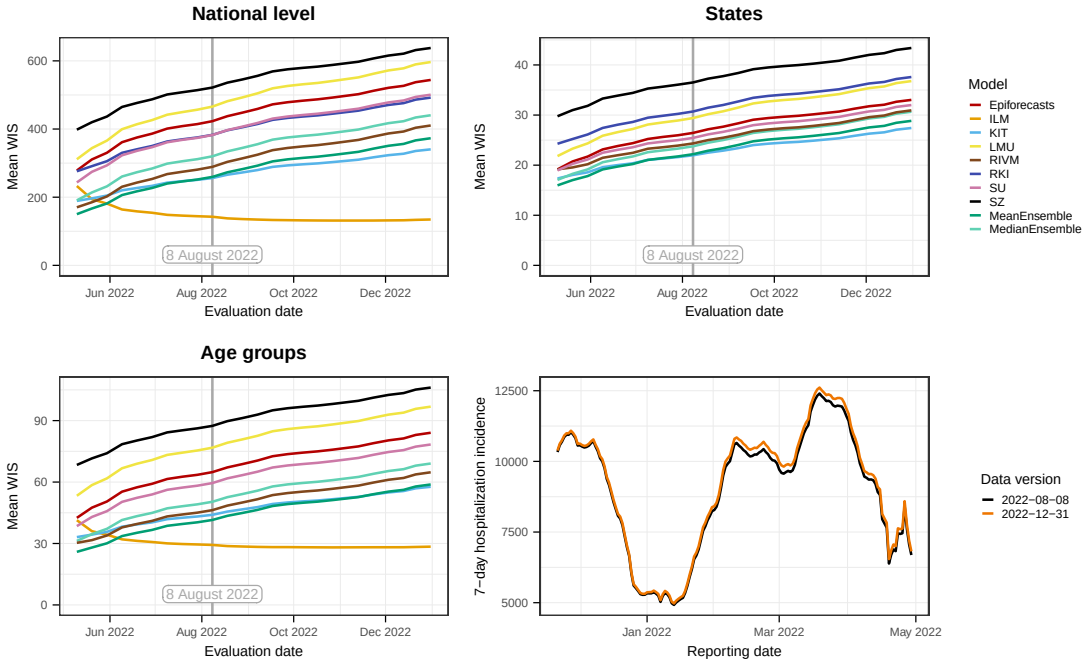


Figure 2.12.: Sensitivity of the scores to the chosen "final" data. Shown is the mean WIS computed with different data versions as the target. The version prespecified in the study protocol is 8 August 2022, marked by a vertical line. Top left: national level. Top right: averaged over states. Bottom left: averaged over age groups. The bottom right panel overlays the national-level data as of 8 August and 31 December to illustrate the importance of late revisions.

maximum delay D. The nowcast target z_t then becomes

$$z_t = \sum_{i=0}^{6} \sum_{d=0}^{D} x_{t-i,d}.$$

In this case, the data for each reference date has the same amount of time to be revised. See Figure A.3 in the Appendix for an illustration of these differently defined target time series. Figure 2.13 shows the results this approach yields for a maximum delay of 40 days, which corresponds roughly to the maximum delay used by most teams. As this target definition is better aligned with the practical implementations teams chose, it is not a surprise that the mean WIS values are lower and coverage is higher. The `ILM` model (in its adjusted version with a maximum delay of 42 days) now shows quite

similar performance to the other approaches, with a tendency to overpredict. The score components for the other models are more balanced and the ensemble nowcasts clearly lead the field. Retrospectively, we think that this definition of targets might have been a more coherent and operationally meaningful approach, see Section 2.4 for a discussion.

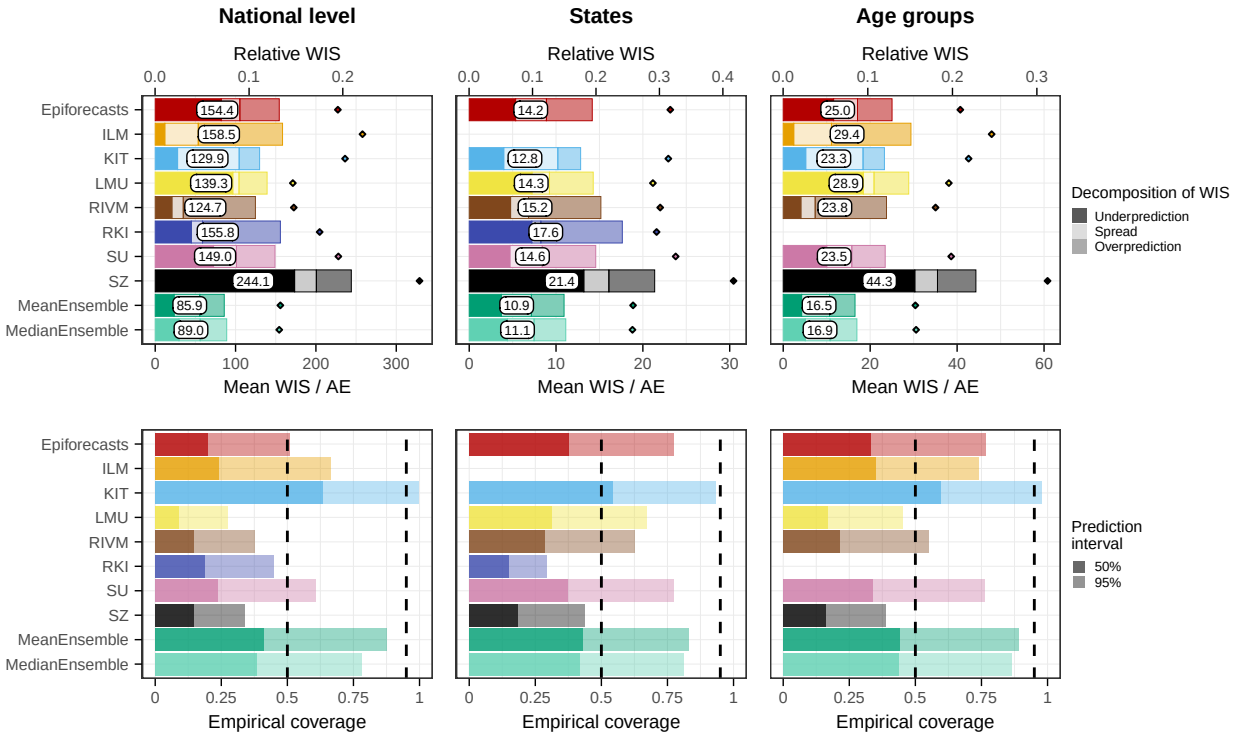


Figure 2.13.: Performance based on the alternative target with a maximum delay of 40 days. Shown are the mean WIS with absolute errors (top) and the empirical coverage (bottom) computed with respect to a revised target defined as the number of hospitalizations reported with a delay of up to 40 days. For the `ILM` model we used the revised model with a maximum delay of 42 days and also recomputed the ensembles with these revised nowcasts. For the other models, the assumed maximum delays are approximately aligned with the redefined target, see Table 2.1.

2.4. Discussion

In this paper, we presented results from a preregistered study to evaluate probabilistic real-time nowcasts of the 7-day hospitalization incidence in Germany from November 2021 to April 2022. We found that all models were able to correct for a large part of the biases caused by reporting delays. Further, we identified calibration of uncertainty intervals as a major challenge, as the empirical coverage rates achieved by most models were considerably below the respective nominal levels. Reasons for insufficient coverage likely include too inflexible modeling of dispersion and delay distributions, and also the fact that most teams truncated delay distributions at a too short maximum delay. The exception was the `ILM` model which also stood out in terms of score-based performance for the national level and across age groups, and to a lesser degree the `KIT` model. These two models incorporated uncertainty quantification using past real-time nowcast errors which proved advantageous when compared to the standard model-based uncertainty intervals.

Our analyses from Sections 2.3.6 and 2.3.7 suggest that the success of the `ILM` model arose from the interplay of two aspects. On the one hand, it used a maximum delay that is longer than those of the other models, but, judging by Figure 2.3, still somewhat too short. On the other hand, the model appears to have a tendency to slightly overpredict the number of hospitalizations added up to a given maximum delay. As these two aspects work in different directions, the resulting nowcasts are overall well-aligned with the defined target (data version from 8 August 2022). Whether nowcasts taking into account case incidence inherently perform better needs to be explored in future work. A combination of different data streams may reduce the dependence of nowcasts on certain assumptions, such as the constant completeness of initial reports. The `Epiforecasts` and `SU` models have been extended in this direction. In an application to COVID-19 deaths in Sweden (Bergström et al., 2022), the inclusion of reported cases and intensive care admissions as leading indicators indeed led to improved predictions.

The `MeanEnsemble`, along with the `KIT` model, performed best across federal states (for which the `ILM` model did not provide nowcasts). Also, it showed very good relative performance for horizons -7 to 0 days, as well as for the revised "rolling" target. We therefore conclude that ensemble approaches are a promising avenue in order to improve

disease nowcasting. However, our case study also illustrates a limitation of unweighted ensembles. The ensemble may have been imbalanced in the sense that a majority of its members followed similar strategies and had similar weaknesses (specifically a downward bias due to neglecting very long delays). A weighted ensemble could have capitalized on the strengths of the `ILM` model, which followed a conceptually different approach and could have served as a counterbalance. However, it is not obvious how ensemble members can be assigned weights in real time in a nowcasting setting. This represents an interesting future research area.

A difficulty we encountered in terms of our study design is that results depend on which data version is used as the "final" one (i.e., the values against which nowcasts are evaluated; Section 2.3.7). As the choice of 8 August 2022 was preregistered and known to all participating teams, the prediction task was well-defined, and we stuck to this choice for our main analysis. Nonetheless, this definition, which was based on the assumption that data would be stable after 100 days, turned out not to be ideal in retrospect. In particular, it implies that for the first day of our study period (22 November 2021), retrospective additions could accumulate over 259 days, while for the last (29 April 2022) this was restricted to 100 days. Defining the nowcast target in a "rolling" fashion as explored in Section 2.3.7 might have been a more appropriate choice. This would have been a more clearly defined modeling task, and modelers would not have had to choose a maximum delay for their models themselves.

The question of whether additions should be ignored from a certain maximum delay onward is closely linked to what these additions actually mean and whether they are relevant from a public health perspective. As mentioned in Section 2.2.1, the 7-day hospitalization incidence also contains hospitalizations that are not primarily due to COVID-19. These hospitalizations have been found to represent a considerable fraction (Heinsch and Schmid-Johannsen, 2022). It seems plausible that very long delays are due to large time differences between the positive test and hospital admission, in which case the share of hospitalizations that are not primarily due to COVID-19 may be high. Also, it can be questioned whether hospitalizations a long time after a positive test are relevant for the real-time assessment of healthcare burden. Both aspects strengthen the case for limiting nowcasting to hospitalizations reported up to a carefully chosen maximum delay.

The definition of the *frozen* values used by the Robert Koch Institute when applying legally defined thresholds can be seen as a strong form of discarding delayed hospitalizations. It has the advantage of simplicity and unambiguity, which are required for actionable guidelines in a legal context. After all, it seems difficult to integrate complex statistical methods with many tuning parameters into a binding legal document. An important downside, however, is that the same *frozen* value can mean rather different things at different time points and in different locations, due to differences in initial reporting completeness. We thus argue that outside of purely legal considerations, nowcasts can provide a more thorough picture of current developments.

All nowcasts generated within the presented collaborative project are available in a public repository (see data availability statement). Time-stamped versions of hospitalization data as available at different points in time can be retrieved from the commit history of the repository as well as directly from Robert Koch Institute (Robert Koch Institute, 2023). We hope that this data can be of use as a benchmarking system for future nowcasting methods. In this context, we note, however, that the present paper is a comparison of *nowcasting systems*, which are given by a statistical model, but also various additional analytical choices, in particular the assumed maximum delay and the length of training data used at each time point. These decisions can have a substantial impact on predictive performance (see Section 2.3.6) and are easier to get right in hindsight than in real time. To ensure a fair comparison, it may therefore be reasonable to use the "rolling" target as discussed in Section 2.3.7.

The nowcasts produced for this project were routinely displayed by numerous German-speaking media, including *Die Zeit*, *Neue Zürcher Zeitung* and *Norddeutscher Rundfunk*. While some displays were limited to the point nowcasts (predictive medians), others made the predictive uncertainty clearly visible. This development should be further encouraged by scientists advising the media on the display of epidemiological data and models. In this context, we also note that data journalists were overall hesitant to use the ensemble nowcasts and prioritized individual nowcasts based on methods described in peer-reviewed publications. Interestingly, the best-performing models in our study were the `MeanEnsemble` and the yet unpublished `ILM` approach. However, our analyses show that all compared approaches provided a good qualitative impression of current incidence

trends and levels, and we consider each of them a helpful addition and improvement over showing uncorrected data.

To conclude, we highlight some advantages of the collaborative nowcasting approach adopted in our study. The ensemble nowcast not only showed strong relative performance but was also the most consistently available nowcast, with almost all other models unavailable due to technical problems on some days during the study period. Additionally, our collaborative approach fostered frequent exchange and interaction among modelers via bi-weekly coordination calls, creating a valuable platform for knowledge sharing, feedback, and collaboration on methodological advancements. Through these interactions, the project facilitated model improvements, as seen for the `LMU`, `RKI` and `KIT` approaches in Section 2.3.6, and fostered discussion on new methodological topics beyond the scope of the present article. For example, the *Epinowcast community* (`https://www.epinowcast.org/`) was established to build and assess real-time analysis tools, publically available in the R package `epinowcast` (Abbott et al., 2021). The benefits of our collaborative approach demonstrate the importance of ongoing communication and cooperation in the development and refinement of epidemiological models, particularly during rapidly evolving public health crises such as infectious disease outbreaks.

Data availability

The nowcasts collected for this study are available in a GitHub repository (`https://github.com/KITmetricslab/hospitalization-nowcast-hub`), with a stable release published at `https://zenodo.org/record/7828604`. The repository also contains the truth data used for evaluation. An interactive dashboard to visualize the nowcasts is provided at `https://covid19nowcasthub.de/`.

Code availability

Code to reproduce results and figures are provided at `https://github.com/dwolffram/hospitalization-nowcast-hub-evaluation`. A list of the participants' code repositories can be found in Appendix A.3.

Appendix A

A.1. Supplementary figures

Figure A.1.: Temporal development of the reporting completeness in the 16 German states. Shown is the reported fraction of the final 7-day hospitalization incidences 0 to 70 days after the respective reference date. Abbreviations of federal states: BB = Brandenburg, BE = Berlin, BW = Baden-Württemberg, BY = Bavaria, HB = Bremen, HE = Hessen, HH = Hamburg, MV = Mecklenburg-Vorpommern, NI = Lower Saxony, NW = North Rhine-Westphalie, RP = Rhineland Pallatinate, SH = Schleswig Holstein, SL = Saarland, SN = Saxony, ST = Saxony Anhalt, TH = Thuringia.

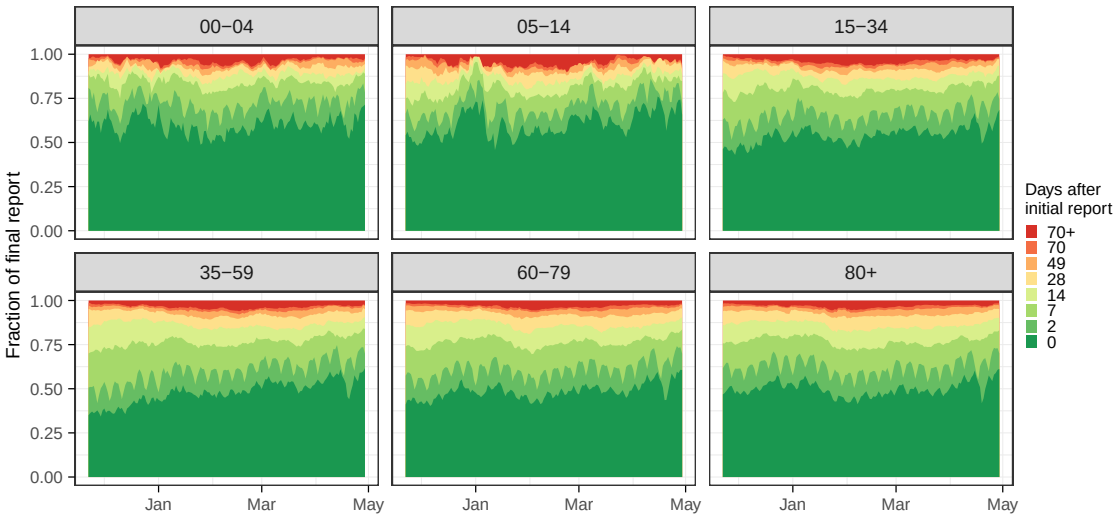


Figure A.2.: Temporal development of the reporting completeness in different age groups. Shown is the reported fraction of the final 7-day hospitalization incidences 0 to 70 days after the respective reference date.

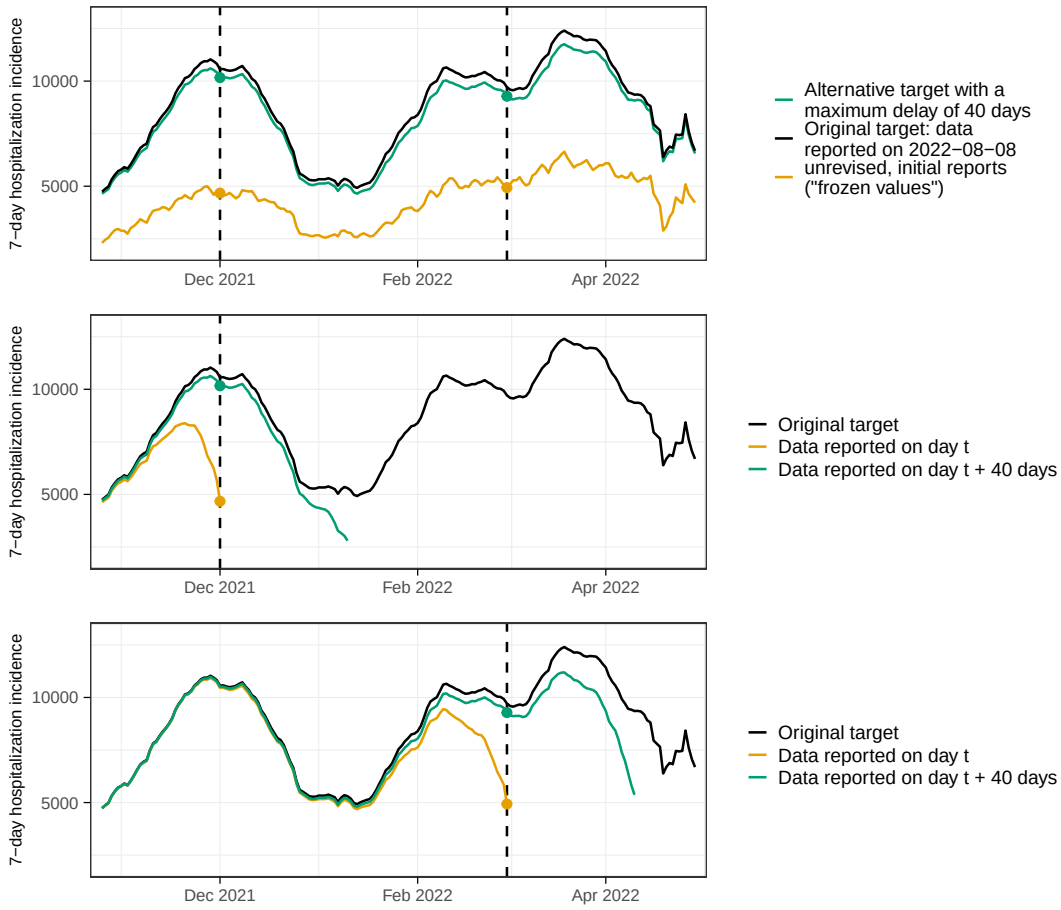


Figure A.3.: Schematic illustration of the alternative target with a maximum delay of 40 days. This target (green) was defined in Section 2.3.7 and is compared to the original target (black) and the frozen values (yellow). In the new target, hospitalizations are only counted if they happen within 40 days after the case report, while in the original target, there was no upper limit on delays. The top panel shows the final version of each time series. The two bottom panels show how they arise from real-time data.

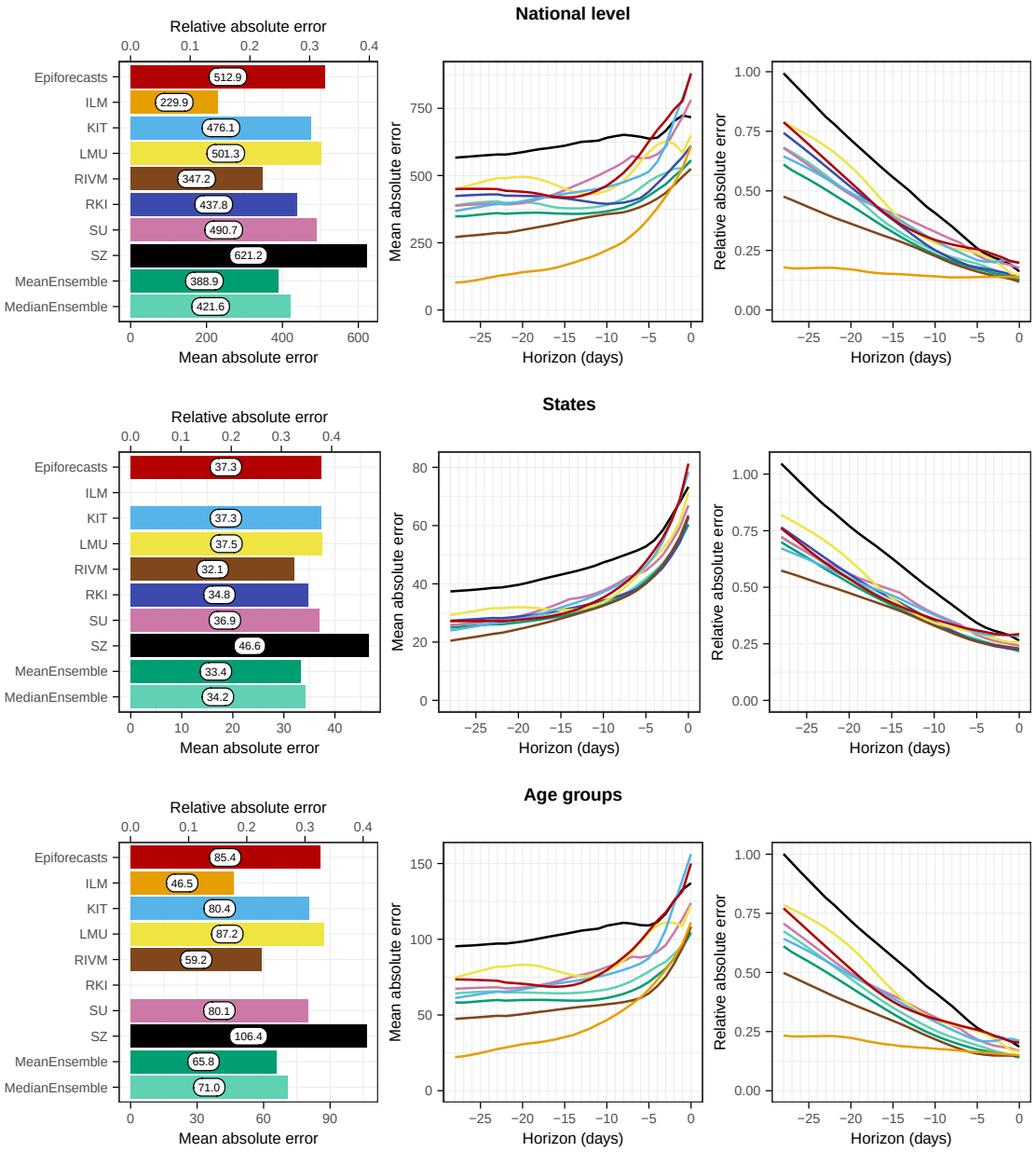


Figure A.4.: Performance of the point predictions (predictive medians). Shown are the mean absolute errors for the national level (top) and averaged across states (middle) and age groups (bottom). The first panel in each row displays the average across all horizons (on the absolute and relative scales). The second and third panels in each row show the mean absolute error and the relative absolute error, respectively, stratified by horizon.

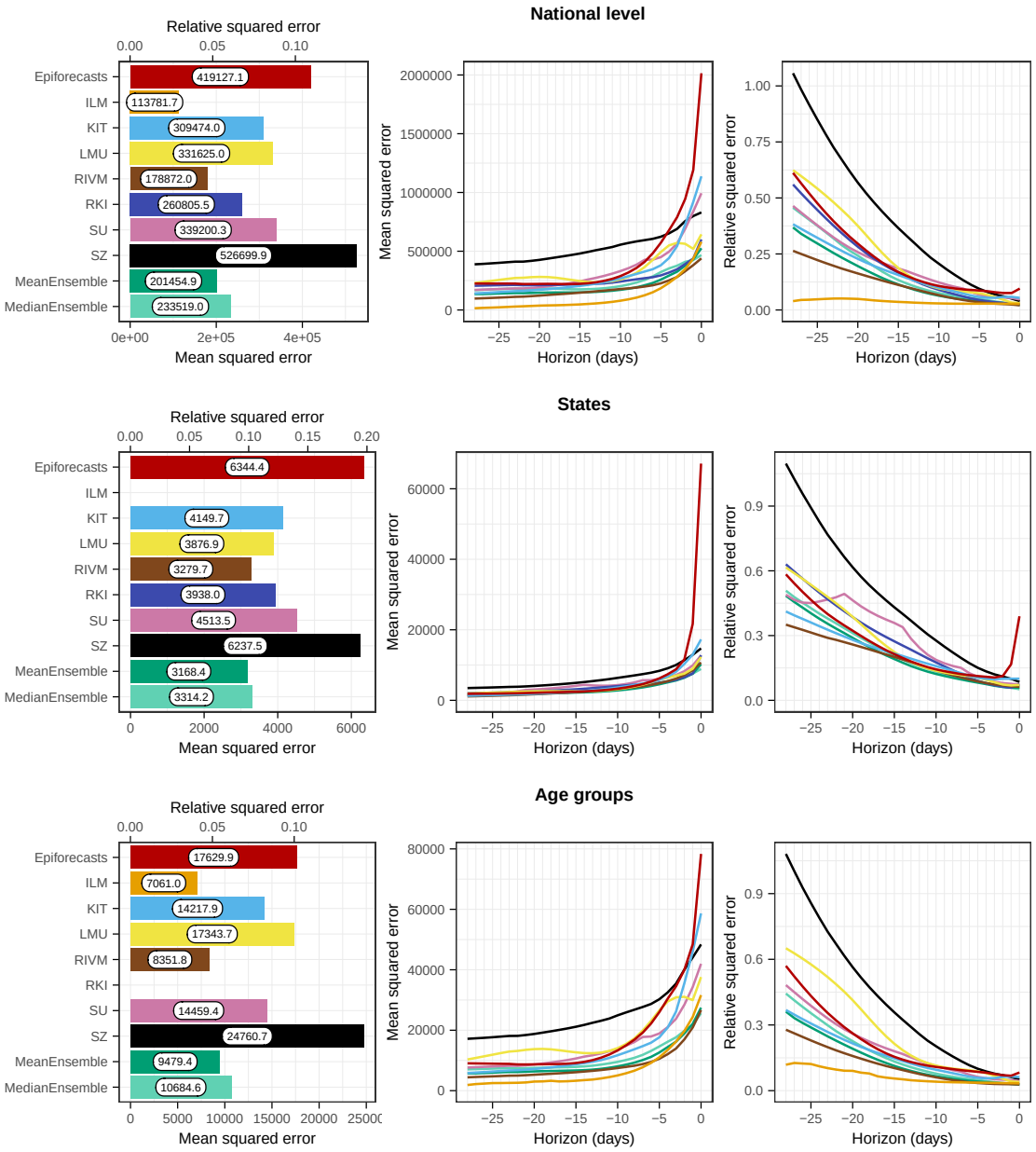


Figure A.5.: Performance of the point predictions (expected values). Shown are the mean squared errors for the national level (top) and averaged across states (middle) and age groups (bottom). The first panel in each row displays the average across all horizons (on the absolute and relative scales). The second and third panels in each row show the MSE and the relative MSE, respectively, stratified by horizon.

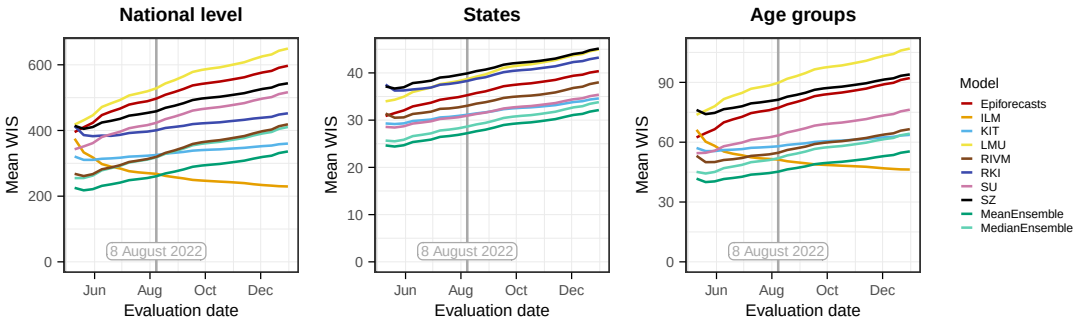


Figure A.6.: Performance on short horizons based on the chosen "final" data. Shown is the mean WIS across horizons from 0-7 days computed using different data versions as the "final" version. The version prespecified in the study protocol is 8 August 2022, marked by a vertical line.

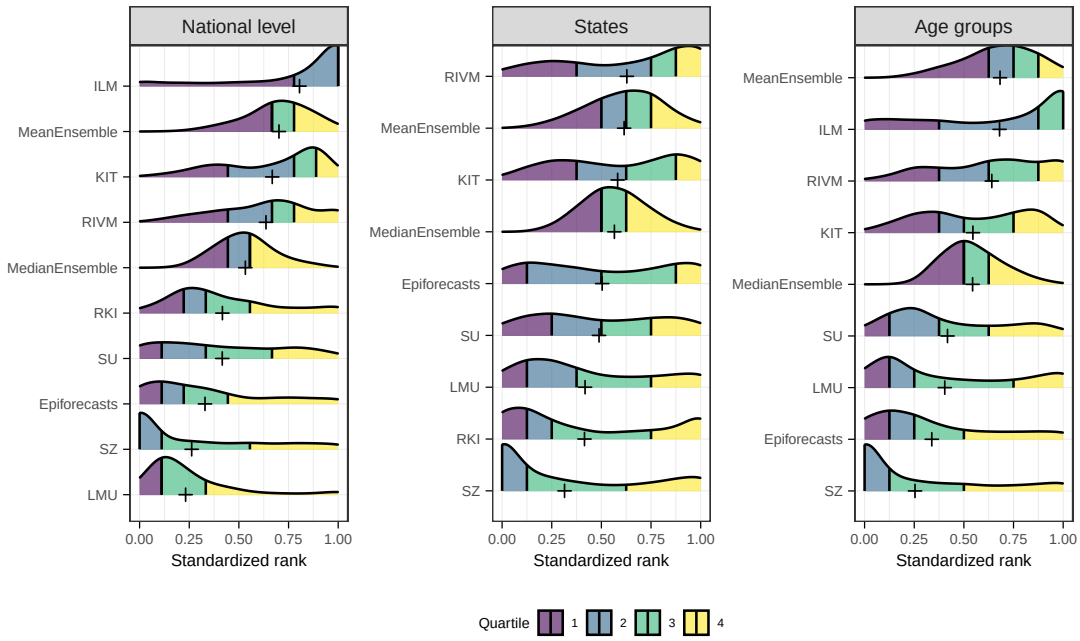


Figure A.7.: Model rank distributions. Distribution of each model's standardized rank for each nowcast-observation pair (see Cramer et al. (2022b) for details on the definition) The models are ordered by the mean standardized rank, which is indicated by a plus sign.

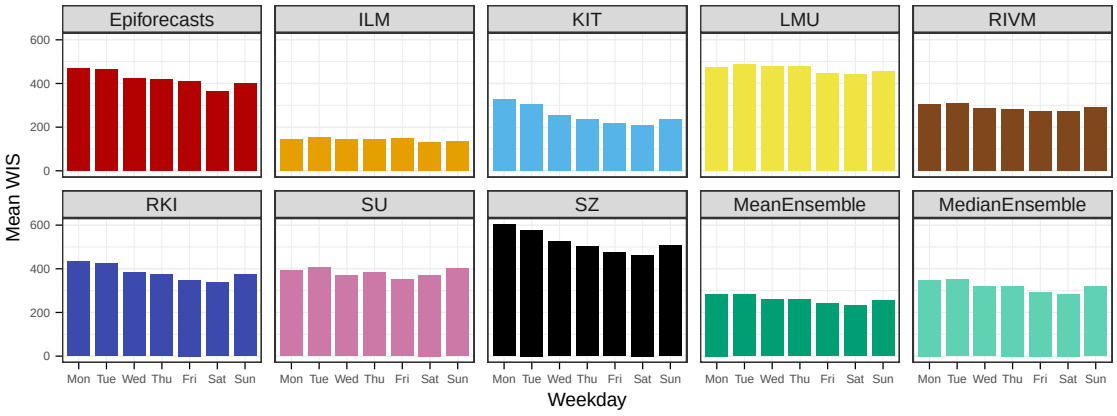


Figure A.8.: Impact of weekday effects on the scores. Mean WIS of same-day nowcasts (with a horizon of 0 days) averaged by weekday.

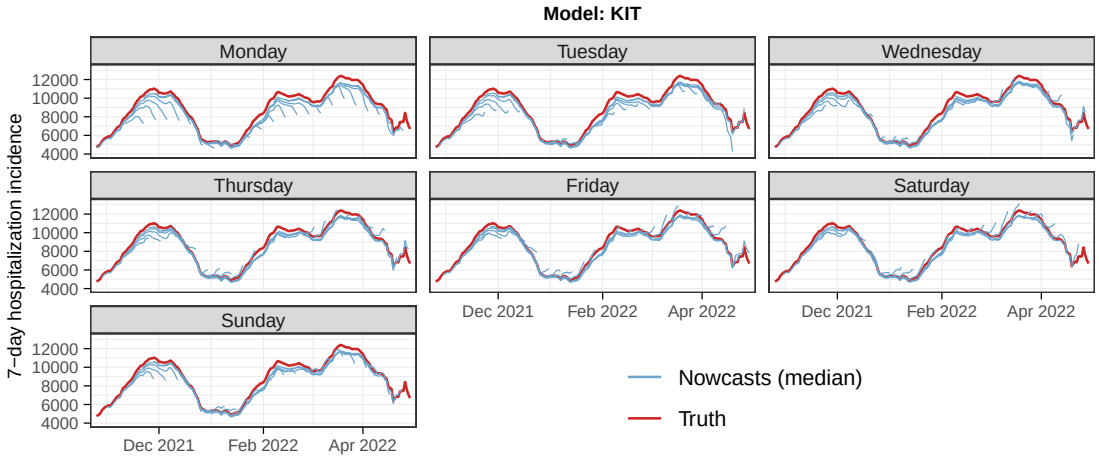


Figure A.9.: Nowcasts of the KIT model issued on different weekdays (0 to 28 days back). Nowcasts issued on Mondays show strong downward biases, while on Saturdays both under- and overprediction occur.

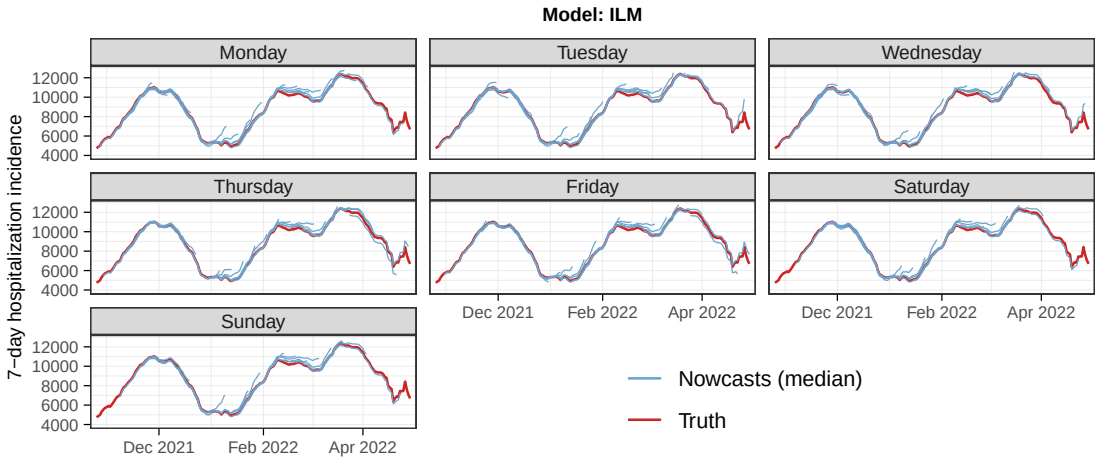


Figure A.10.: Nowcasts of the ILM model issued on different weekdays (0 to 28 days back). No clear weekday patterns can be discerned.

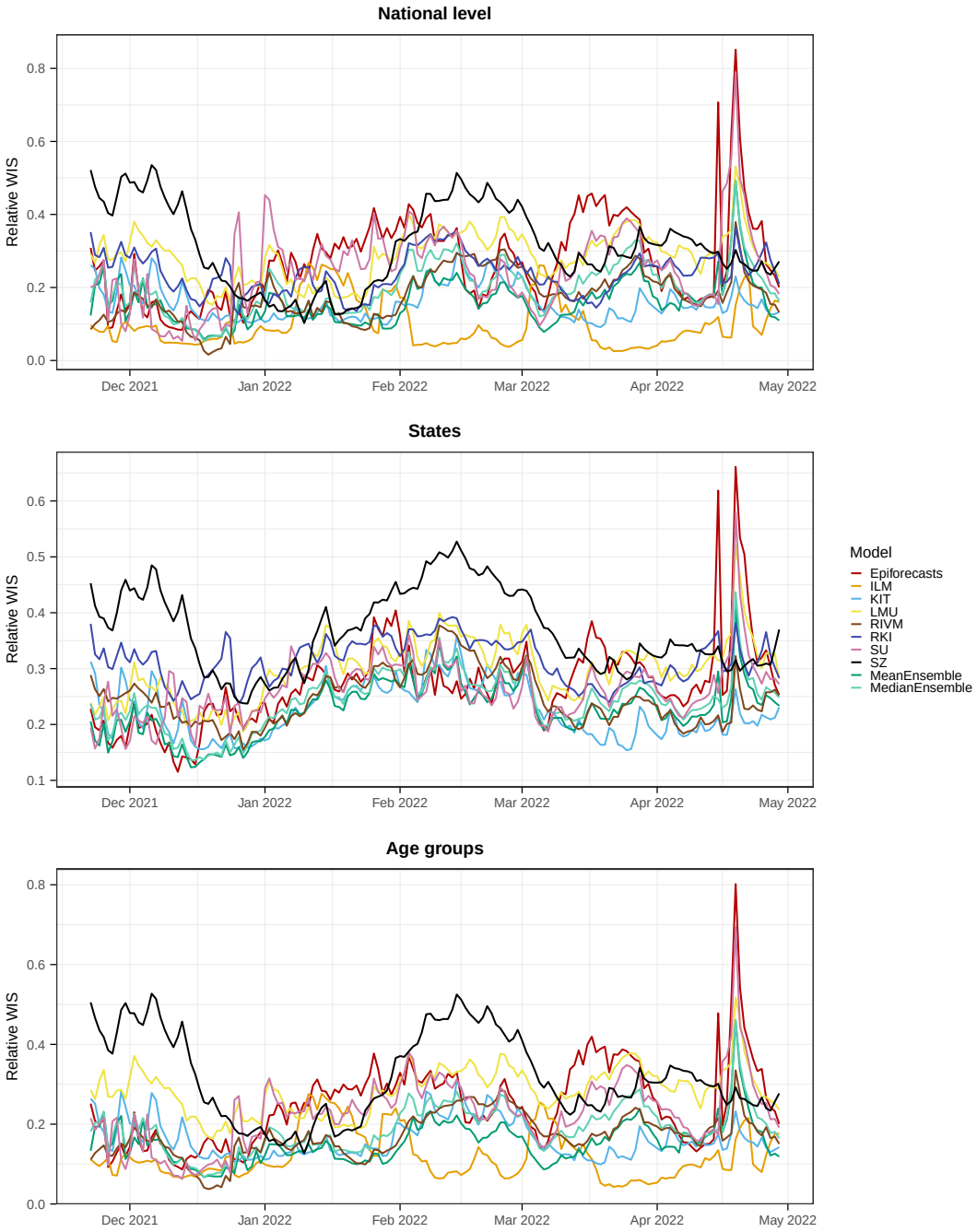


Figure A.11.: Performance over time. Relative mean WIS across all horizons by nowcast date and stratification level. Top: national level; middle: across states; bottom: across age groups.

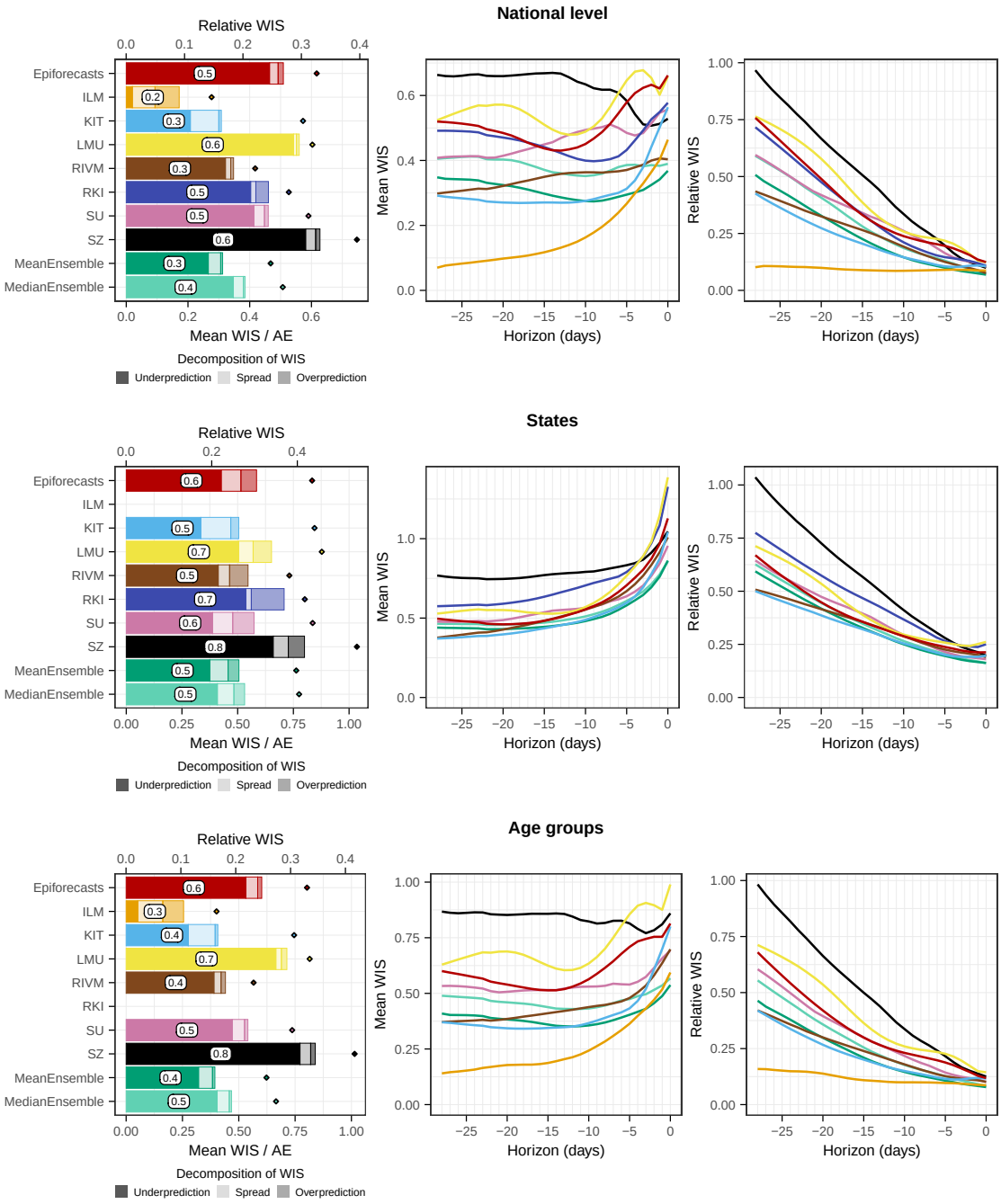


Figure A.12.: Scores computed after standardization by population. Shown are the mean WIS and AE as in Figure 2.7 but for hospitalizations by 100,000 inhabitants. This gives similar weight to all states irrespective of their population size.

A.2. Deviations from study protocol and completeness of nowcasts

As we have deviated in some minor parts from the study protocol, we provide a list of these adjustments.

Table A.1.: Deviations from the study protocol.

Description	Reasoning
Inclusion of weekends into the study period.	Initially, weekends were excluded from the study period as we expected submission to require some human intervention at least initially. As all teams completely automated their procedures very quickly, we were able to update models on weekends and include them in the evaluation.
Inclusion of the Christmas period into the study period.	In the study protocol, the Christmas period was excluded from the study period as (i) we expected irregular reporting behavior and (ii) we did not want to oblige modelers to perform any manual steps for submission during this period. However, as there were no unusual patterns and all submissions were completely automated, we decided to include this period. We note that, contrary to our expectations, the Easter period showed some unusual reporting patterns. We nonetheless kept it in the evaluation as removing periods for which nowcasts were expected to work normally but did not would unduly embellish our results.
Omission of the retrospective study period from 1 July 2021 to 19 November 2021.	We initially planned to include a retrospective study period in order to contrast the performance of methods in retrospect and in real time. However, as only two teams provided retrospective nowcast before the start of the study period, we chose to omit this aspect. Instead, we chose to include an analysis of four revised versions of contributed models applied retrospectively to the period 22 November to 29 April.
Omission of nowcast targets for which even including fill-in nowcasts results could not be obtained from all methods.	For a very small set of targets we were unable to obtain submissions from all models, even including fill-in nowcasts. As these represented a negligible fraction of all targets (0.3%) we pragmatically chose to omit these from the main analysis in order to achieve a balanced data set.
Omission of interval coverage results at the 80% level.	As interval coverage results at the 80% level provided no additional insights and led to overly full figures we chose to omit them.
Definition of naïve baseline model.	At the time of writing the protocol we had decided that a naïve baseline model should be included, but it was unclear how it should be defined. The `FrozenBaseline` used in the main analysis was only defined during the work on the manuscript.
Tightening of ensemble inclusion criteria.	During the study period, we realized that one model (`SZ`) occasionally issued nowcast values below the already known values (which in almost all cases only get corrected upwards). This is due to a smoothing step that is included in the procedure. We decided to exclude these from the ensemble. Specifically, submissions were excluded from the ensemble whenever the median or mean nowcast was below the already known value.

Table A.2.: Missingness of real-time submissions by the participating teams. Apart from the targets listed in Table A.3, all of these could be imputed with fill-in nowcasts.

Model	Dates without submission
`Epiforecasts`	25–26 Jan 2022
`ILM`	27–28 Nov 2021, 24, 26, 30 Dec 2021, 16 Jan 2022, 8–20 April 2022
`RIVM`	8 Dec 2021, 23 Apr 2022
`RKI`	8 Dec 2021; all 0 and -1 day nowcasts
`SU`	27–28 Nov 2021, 5 Dec 2022

Table A.3.: Nowcast targets for which no complete sets of submissions could be obtained. These amount to 394 nowcast targets among the 109,968 considered in total.

Dates	Excluded targets	Reason
22 November 2021	Horizons -1 and 0 days, all strata	On the first day of our study several models provided only nowcasts from -2 days backward.
22–24 November 2021	Horizons -23 to -28 days, all strata.	The `SZ` model initially only provided nowcasts three weeks back.
31 Jan, 1 Feb 2022	State of Hamburg, all horizons.	Nowcasts from `Epinowcasts` model not available due to convergence issues.

Table A.4.: Other decisions in response to unexpected difficulties.

Description	Reasoning
Imputation of 0.1-quantiles from `Epiforecasts` model via an interpolation/normal approximation.	For several months the `Epiforecasts` model did not provide 0.1 quantiles, which was only noticed towards the end of the study period. in order to be able to evaluate the WIS without having to rerun the model for all concerned dates, we imputed the 0.1-quantile by interpolating between the 0.025 and the 0.25 quantile based on a normal approximation (which implies that the 0.1 quantile is almost exactly halfway between the 0.025 and 0.25 quantiles).
`LMU` nowcasts for Saarland and Bremen were replaced by nowcasts from the retrospectively revised model version discussed in Section 2.3.6.	In their real-time submissions, the `LMU` team only reported point nowcasts for the states of Saarland and Bremen, which are considerably smaller than the other states (1M and 700k inhabitants, respectively). To be able to nonetheless evaluate the WIS and keep these two states in the overall evaluation, we used the revised nowcasts as discussed in Section 2.3.6 as these contained all quantiles for these states. We consider this defensible as the role of these two states in the overall evaluation under WIS is very small (the WIS typically scales with the order of magnitude of the target).
Filling in some missing entries of `ILM` with nowcasts from the updated model.	Nowcasts from 22-26 and 29 November 2021 were missing entries for horizons -28, -1, and 0 days. To fill these in, we used the revised nowcasts as discussed in Section 2.3.6.
Removal of a small number of obviously erroneous nowcasts for the `RKI` model.	In a handful of instances, the `RKI` model submitted obviously erroneous nowcasts for the 0-day horizon. These stated values of more than 1 billion hospitalizations. We replaced these with the respective -1 day nowcasts.

A.3. Repositories of participating teams

- `Epiforecasts`: https://epiforecasts.io/eval-germany-sp-nowcasting/
- `ILM`: https://github.com/Stochastik-TU-Ilmenau/ILM-prop
- `KIT`: https://github.com/KITmetricslab/hospitalization-nowcast-hub/tree/main/code/baseline
- `LMU`: https://github.com/MaxWeigert/Nowcasting_covid19_hospitalizations

- RIVM: https://github.com/kassteele/Nowcast-hub

- SU: https://github.com/FelixGuenther/hospitalization-nowcast-hub_SU-public

A.4. Sensitivity analysis via pairwise comparisons

Motivation and procedure

In some instances, teams failed to submit nowcasts in time and had to fill them in post hoc. Allowing them to do so may seem lenient as, in principle, teams could use additional information, thus unfairly improving their nowcasts. As specified in our protocol, we thus perform a sensitivity analysis based purely on nowcasts submitted on time of the respective day.

As each team failed to submit nowcasts on different days this leads to a setting where methods need to be compared based on incongruent sets of nowcasting tasks. In this setting, the relative WIS could still be evaluated for each method by considering only the subset of targets treated by the respective method. This, however, ignores that improving upon the naïve baseline is easier for certain locations, age groups, and time periods than others. To handle this difficulty, we use the *pairwise comparison* approach suggested in Cramer et al. (2022b). It is based on the assumption that achieving good nowcast performance *relative to all other considered methods* is similarly difficult across locations, age groups, and time periods. Considering a set of N models (including the baseline model), the relative WIS corrected for missing submissions is determined as follows:

1. In the first step for each pair of models i, j we compute the ratio

$$\theta_{ij} = \frac{\text{mean WIS achieved by model } i}{\text{mean WIS achieved by model } j}.$$

2. For each model i we then compute the geometric average of the ratios θ_{ij} achieved in the comparisons to all other models

$$\theta_i = \left(\prod_{i=1}^{N} \theta_{ij} \right)^{1/N}.$$

3. Lastly, we re-scale the θ_i to the one achieved by the baseline model to obtain the relative WIS:

$$\text{adjusted relative WIS of model } i = \frac{\theta_i}{\theta_{\text{BL}}},$$

where θ_{BL} refers to the baseline model (`FrozenBaseline`).

If all models submitted all required nowcasts it is straightforward to show that this boils down to the regular relative WIS as defined in Section 2.2.5. If some submissions are missing for certain models, the procedure will adjust the relative WIS to how well other models fared on the respective subset of addressed targets.

Results

Table A.5 compares the relative WIS computed using fill-in nowcasts as in the main analysis and the pairwise comparison approach. The differences are very modest, meaning that the missingness of nowcasts does not substantially affect the results. This is not surprising, given the low number of missing submissions.

Table A.5.: Comparison of relative WIS values obtained using retrospective fill-in nowcasts and the pairwise comparison approach from Cramer et al. (2022b) (PC).

	National level		States		Age groups	
	PC	fill-in	PC	fill-in	PC	fill-in
`Epiforecasts`	0.2679	0.2690	0.2686	0.2686	0.2465	0.2472
`FrozenBaseline`	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
`ILM`	0.0907	0.0907	–[1]	–[1]	0.1124	0.1117
`KIT`	0.1627	0.1628	0.2227	0.2232	0.1676	0.1677
`LMU`	0.2961	0.2962	0.2993	0.2988	0.2925	0.2926
`MeanEnsemble`	0.1643	0.1651	0.2254	0.2254	0.1574	0.1581
`MedianEnsemble`	0.2034	0.2034	0.2414	0.2415	0.1919	0.1920
`RIVM`	0.1845	0.1838	0.2471	0.2473	0.1767	0.1764
`RKI`	–[2]	0.2435	–[2]	0.3120	–[1]	–[1]
`SU`	0.2427	0.2432	0.2590	0.2585	0.2265	0.2268
`SZ`	0.3314	0.3317	0.3703	0.3709	0.3330	0.3332

[1] No nowcasts submitted for this target.

[2] WIS could only be evaluated for fill-in nowcasts as real-time submissions did not contain all required quantiles.

A.5. Documentation of the KIT model

As the KIT model was conceived as a conceptually simple (though not naïve) reference model for the current study we provide a brief documentation of its methodology.

Notation Denote by $X_{t,d}, d = 0, \ldots, D$ the number of hospitalizations for reference date t which appear in the data set at day $t + d$ and by

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

the number of hospitalizations reported for reference date t up to day $t + d$. Moreover, denote by

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

the total number of reported hospitalizations for t, where D denotes an assumed maximum possible delay. In the following, we denote by X_t, etc. a random variable and by x_t the corresponding observation.

The observed $x_{t,d}$ as available at a given time point t^* can be arranged into the so-called *reporting triangle*, see Table A.6.

Table A.6.: Illustration of the reporting triangle for time t^* and $D = 5$. Quantities known at time t are shown in black, yet unknown quantities are shown in gray.

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

As we will focus on seven-day hospitalization incidences we moreover need to consider rolling sums over windows of length W (usually $W = 7$)

$$Y_t = \sum_{w=0}^{W-1} X_{t-w}.$$

Goal Our aim is to estimate or *nowcast* Y_t based on the information available at time $t^* \geq t$. We do not take into account any information other than data on hospitalizations and their reporting delays, meaning that we model

$$Y_t \mid X_{s,d} : s + d \leq t^*, d \geq 0.$$

Point nowcast The following describes a simple heuristic to obtain a point prediction of Y_t based on information available at time t^*.

We start by imputing

$$x_{t^*,1} = x_{t^*,0} \times \frac{\sum_{i=1}^{t^*-1} x_{t^*-i,1}}{\sum_{i=1}^{t^*-1} x_{t^*-i,0}},$$

i.e. use a simple multiplication factor computed from the complete rows of our data set. Next, we compute

$$x_{t^*,2} = x_{t^*,\leq 1} \times \frac{\sum_{i=1}^{t^*-1} x_{t^*-i,2}}{\sum_{i=1}^{t^*-1} x_{t^*-i,\leq 1}},$$

where in the computation of

$$x_{t^*-i,\leq 1} = x_{t^*-i,\leq 0} + x_{t^*-i,1}$$

we just treat the $x_{t^*-i,1}$ imputed in the first step as if it was a known value. The same can be done for

$$x_{t^*-1,2} = x_{t^*-1,\leq 1} \times \frac{\sum_{i=1}^{t^*-1} x_{t^*-i,2}}{\sum_{i=1}^{t^*-1} x_{t^*-i,\leq 1}}.$$

We repeat this same procedure to fill in the missing values of the reporting triangle step by step, moving from the left to the right and the bottom to the top.

This is equivalent to the following slightly more formal formulation: We denote by π_d the probability that a hospitalization with reference date t appears in the data on day

$t + d$ and by

$$\pi_{\leq d} = \sum_{i=0}^{d} \pi_i$$

the probability that such a hospitalization appears in the data no later than $t + d$. We introduce

$$\theta_d = \frac{\pi_d}{\pi_{\leq d-1}},$$

which allows us to formulate the recursion

$$\pi_{\leq d} = (1 + \theta_d)\pi_{\leq d-1}.$$

To estimate the θ_d for $d = 1, \ldots, D < t$ based on quantities available at time t^* we use

$$\hat{\theta}_d(t^*) = \frac{\sum_{j=d}^{J} X_{t^*-j,d}}{\sum_{j=d}^{J} X_{t^*-j,\leq d-1}},$$

where J is the number of past observations to include in the estimation (in practice it is often helpful to use only a recent subset rather than the entire available history). Note that we treat this estimate as a function of t^* as it may change over time. Estimates of the probabilities $\pi_{\leq d}$ can then be obtained as

$$\hat{\pi}_{\leq d}(t^*) = (1 + \hat{\theta}_d)\hat{\pi}_{\leq d-1}.$$

These can subsequently serve to estimate the total number X_t of hospitalizations with reference date t based on the $X_{t,\leq t^*-t}$ hospitalizations already reported by time t^*:

$$\hat{X}_t(t^*) = \frac{X_{t,\leq t^*-t}}{\hat{\pi}_{\leq t^*-t}(t^*)}.$$

We can also compute the estimates for the respective number of hospitalizations reported with a given delay $d > t^* - t$, which is given by

$$\hat{X}_{t,d}(t^*) = \hat{\pi}_d(t^*)\hat{X}_t(t^*).$$

In the last step we move to the rolling sum Y_t, which we estimate as

$$\hat{Y}_t(t^*) = \sum_{w=0}^{W-1} \hat{X}_{t-w}(t^*).$$

Uncertainty quantification Our general idea to quantify the nowcast uncertainty for $\hat{Y}_t(t^*)$ is to generate point nowcasts $\hat{Y}_{t-1}(t^*-1), \hat{Y}_{t-2}(t^*-2), \ldots, \hat{Y}_{t-K}(t^*-K)$ for $K > D$ past time points, each based on the information available at the respective time point. These could then be compared to the corresponding observations $Y_{t^*-1}, \ldots, Y_{t^*-K}$, and nowcast dispersion could be based on a simple parametric model. However, two aspects need to be taken into account:

- The information available at t^*, on which the nowcast $\hat{Y}_t(t^*)$ is based, already implies a lower bound for Y_t, namely the hospitalizations which have already been observed. Only the hospitalizations for reference date t which will be reported after t^* need to be modeled probabilistically. We thus introduce the decomposition

  $$Y_t = Y_{t,\leq t^*-t} + Y_{t,>t^*-t}.$$

  Here,

  $$Y_{t,\leq t^*-t} = \sum_{w=0}^{W-1} \sum_{d=0}^{D} X_{t-w,d} \times \mathbb{I}(-w+d \leq t^*-t)$$

  are those already observed by t^* (i.e., the lower bound) and

  $$Y_{t,>t^*-t} = \sum_{w=0}^{W-1} \sum_{d=0}^{D} X_{t-w,d} \times \mathbb{I}(-w+d > t^*-t)$$

  are those yet to be observed. We only need to quantify the uncertainty about the latter.

- At time t^*, the realizations of $Y_{t,>t^*-t}$ are only available for $t \leq t^* - D$. If we only want to use complete observations we would need to discard a lot of recent information.

We therefore construct a set of observations $Z_{t-j,>t^*-t}, j = 1, \ldots, K$ and corresponding point predictions $\hat{Z}_{t-j,>t^*-t}, (t^*-j)$ as follows:

- For $j = D, \ldots, K$ we can simply set

$$Z_{t-j,>t^*-t} = Y_{t-j,>t^*-t}$$

and point predictions $\hat{Z}_{t-j,>t^*-t}(t^* - j) = \hat{Y}_{t-j,>t^*-t}(t^* - j)$ as all relevant information are already available at t^*.

- For $j = 1, \ldots, D-1$ we use partial observations

$$\begin{aligned} Z_{t-j,>t^*-t} &= \sum_{w=0}^{W-1} \sum_{d=0}^{D} X_{t-j-w,d} \times \mathbb{I}(\underbrace{t-j-w+d \leq t^*}_{\text{"already observed at } t^*\text{"}}), \\ &= Y_{t-j,>t^*-t} - Y_{t-j,>t^*-t+j} \end{aligned}$$

which are restricted to hospitalizations already reported by time t^*, so that $Z_{t-j,>t^*-t}$ can be evaluated. The corresponding point nowcasts are given by

$$\hat{Z}_{t-j,>t^*-t} = \sum_{w=0}^{W-1} \sum_{d=0}^{D} \hat{X}_{t-j-w,d} \times \mathbb{I}(\underbrace{t-j-w+d \leq t^*}_{\text{"already observed at } t^*\text{"}}).$$

We then pragmatically assume that

$$Z_{t-j} \mid \hat{Z}_{t-j}(t^* - j) \sim \text{NegBin}(\text{mean} = \hat{Z}_{t-j}(t^* - j), \text{disp} = \psi_{t^*-t}),$$

where we parameterize the negative binomial distribution via its mean and the dispersion (size) parameter ψ_{t^*-t}. Note that the dispersion parameter depends on how far back into the past we nowcast (i.e., how much information has already accumulated between $t - j$ and $t^* - j$). The parameters $\psi_0, \ldots, \psi_D$ are then estimated via maximum likelihood. To avoid issues with zero expectations we add 0.1 to the expected values when feeding them into the maximum likelihood procedure.

The predictive distributions for Y_t are then set to $\text{NegBin}(\text{mean} = \hat{Y}_{t,>t^*-t}(t^*), \text{size} = \psi_{t^*-t})$, shifted by $Y_{t,\leq t^*-t}$. As a motivation for the use of partial observations in the

estimation of the overdispersion parameters, we note that if

$$A \sim \text{NegBin}(\text{mean} = \hat{A}, \text{disp} = \psi)$$

and

$$B \mid A \sim \text{Bin}(A, \pi)$$

one gets

$$B \sim \text{NegBin}(\text{mean} = \pi\hat{A}, \text{disp} = \psi).$$

The negative binomial distribution with a given dispersion parameter is thus closed to binomial subsampling, with only the expectation, but not the size parameter changing. It is therefore defendable to assume the same size parameter for the constructed partial observations $Z_{t-j,>t^*-t}$ and the actual $Y_{t-j,>t^*-t}$ which we would use if they were already available.

Parameter choices To apply the suggested method, the numbers J and K of past observations are used to estimate the nowcast mean and dispersion parameters. Here one needs to strike a balance between a sufficient amount and recency of training data. We set both J and K to 60 days without further assessing the impact on nowcast quality. The maximum delay of D was set to 40 days.

3. A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave

3.1. Introduction

Forecasting is one of the key purposes of epidemic modelling, and despite being related to the understanding of underlying mechanisms, it is a conceptually distinct task (Keeling and Rohani, 2008; Baker et al., 2018). Explanatory models are often strongly idealized and tailored to specific settings, aiming to shed light on latent biological or social mechanisms. Forecast models, on the other hand, have a strong focus on observable quantities, aiming for quantitatively accurate predictions in a wide range of situations. While understanding of mechanisms can provide guidance to this end, forecast models may also be purely data-driven. Accurate disease forecasts can improve the situational awareness of decision makers and facilitate tasks such as resource allocation or planning of vaccine trials (Dean et al., 2020). During the COVID-19 pandemic, there has been a surge in research activity on epidemic forecasting. Contributions vary greatly in terms of purpose, forecast targets, methods, and evaluation criteria. An important distinction is between longer-term scenario or what-if projections and short-term forecasts (Reich and Rivers, 2020). The former attempt to discern the consequences of hypothetical scenarios (e.g., intervention strategies), a task closely linked to causal statements as made by explanatory models. Scenarios typically remain counterfactuals and thus cannot be evaluated directly using subsequently observed data. Short-term forecasts, on the other

hand, refer to brief time horizons, at which the predicted quantities are expected to be largely unaffected by yet unknown changes in public health interventions. This makes them particularly suitable to assess the predictive power of computational models, a need repeatedly expressed during the pandemic (Nature Publishing Group, 2020).

Rigorous assessment of forecasting methods should follow several key principles. Firstly, forecasts should be made in real time, as retrospective forecasting often leads to overly optimistic conclusions about performance. Real-time forecasting poses many challenges (Desai et al., 2019), including noisy or delayed data, incomplete knowledge on testing and interventions as well as time pressure. Even if these are mimicked in retrospective studies, some benefit of hindsight remains. Secondly, in a pandemic situation with low predictability, forecast uncertainty needs to be quantified explicitly (Held et al., 2017; Funk et al., 2019). Lastly, forecast studies are most informative if they involve comparisons between multiple independently run methods (Viboud and Vespignani, 2019). Such collaborative efforts have led to important advances in short-term disease forecasting prior to the pandemic (Viboud et al., 2018; Del Valle et al., 2018; Johansson et al., 2019; Reich et al., 2019a). Notably, they have provided evidence that ensemble forecasts combining various independent predictions can lead to improved performance, similar to what has been observed in weather prediction (Gneiting and Raftery, 2005).

The German and Polish COVID-19 Forecast Hub is a collaborative project which, guided by the above principles, aims to collect, evaluate, and combine forecasts of weekly COVID-19 cases and deaths in the two countries. It is run in close exchange with the US COVID-19 Forecast Hub (Ray et al., 2020; COVID-19 Forecast Hub Team, 2020) and aims for compatibility with the forecasts assembled there. Close links moreover exist to a similar effort in the United Kingdom (Funk et al., 2020). Other conceptually related works on short-term forecasting or baseline projections include those by consortia from Austria (Bicher et al., 2020) and Australia (Golding et al., 2020) as well as the European Centre for Disease Prevention and Control (European Centre for Disease Prevention and Control, 2020a,c, ECDC). In a German context, various nowcasting efforts exist (Günther et al., 2021). All forecasts assembled in the German and Polish COVID-19 Forecast Hub are publicly available (`https://github.com/KITmetricslab/covid19-forecast-hub-de`, German and Polish COVID-19 Forecast Hub Team (2021b)) and can be explored

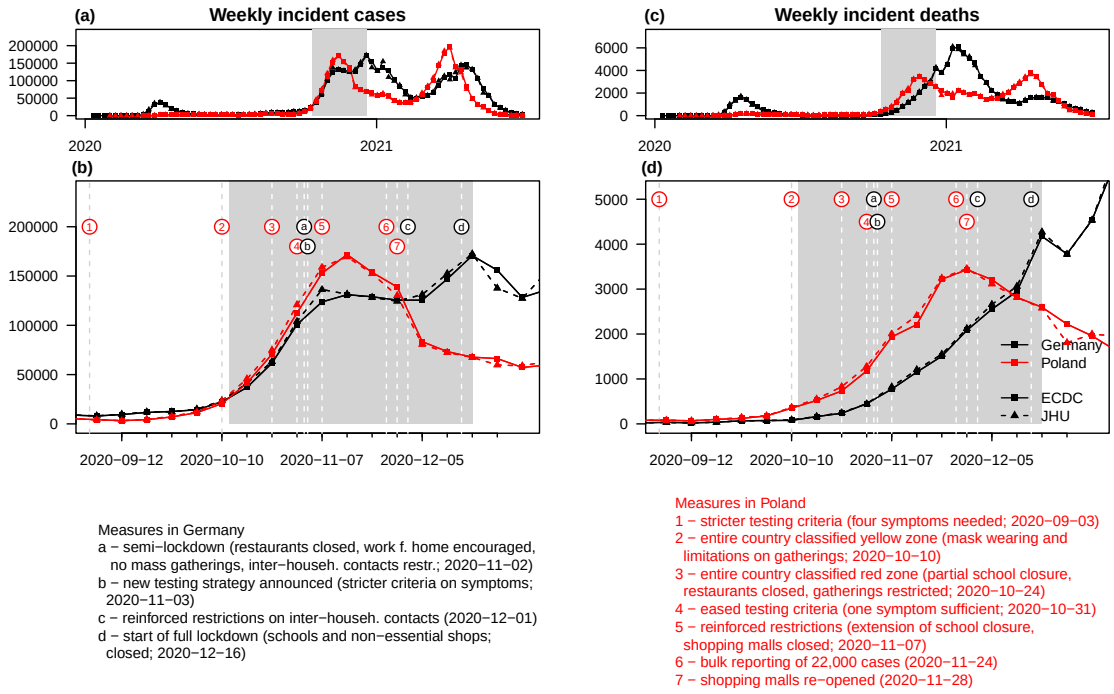


Figure 3.1.: Forecast evaluation period. Weekly incident **a, b** confirmed cases and **c, d** deaths from COVID-19 in Germany and Poland according to data sets from the European Centre for Disease Prevention and Control (ECDC) and the Center for Systems Science and Engineering at Johns Hopkins University (JHU). The study period covered in this paper is highlighted in grey. Important changes in interventions and testing are marked by letters/numbers and dashed vertical lines. Sources containing details on the listed interventions are provided in Appendix B.5.

interactively in a dashboard (https://kitmetricslab.github.io/forecasthub). The Forecast Hub project moreover aims to foster exchange between research teams from Germany, Poland, and beyond. To this end, regular video conferences with presentations on forecast methodologies, discussions, and feedback on performance were organized.

In this work we present results from a prospective evaluation study based on the collected forecasts. The evaluation procedure was prespecified in a study protocol (Bracher et al., 2020) which we deposited at the registry of the Open Science Foundation (OSF) on 8 October 2020. The evaluation period extends from 12 October 2020 (first forecasts issued) to 19 December 2020 (last observations made). This corresponds to the onset of the second wave of the pandemic in both countries. It is marked by strong virus

circulation and changes in intervention measures and testing strategies, see Figure 3.1 for an overview. This makes for a situation in which reliable short-term forecasting is both particularly useful and particularly challenging. Thirteen modelling teams from Germany, Poland, Switzerland, the United Kingdom and the United States contributed forecasts of weekly confirmed cases and deaths. Both targets are addressed on the incidence and cumulative scales and one through four weeks ahead, with evaluation focused on one and two weeks ahead. We find considerable heterogeneity between forecasts from different models and an overall tendency to overconfident forecasting, i.e. lower than nominal coverage of prediction intervals. While for deaths, a number of models were able to outperform a simple baseline forecast up to four weeks into the future, such improvements were limited to shorter horizons for cases. Combined ensemble predictions show good relative performance in particular in terms of interval coverage, but do not clearly dominate single-model predictions. Conclusions from ten weeks of real-time forecasting are necessarily preliminary, but we hope to contribute to an ongoing exchange on best practices in the field. Note that the considered period is the last one to be unaffected by vaccination and caused exclusively by the "original" wild type variant of the virus. Early January marked both the start of vaccination campaigns and the likely introduction of the B.1.1.7 (alpha) variant of concern in both countries. Our study will be followed up until at least March 2021 and may be extended beyond.

3.2. Results

In the following, we provide specific observations made during the evaluation period as well as a formal statistical assessment of performance. Particular attention is given to combined ensemble forecasts. Forecasts refer to data from the European Centre for Disease Prevention and Control (2020b, ECDC) or Johns Hopkins University Center for Systems Science and Engineering (Dong et al., 2020, JHU CSSE); see the Methods section for the exact definition of targets and ensemble methods. Visualizations of one- and two-week-ahead forecasts on the incidence scale are displayed in Figures 3.2 and 3.3, respectively, and will be discussed in the following. These figures are restricted to models submitted over (almost) the entire evaluation period and providing complete forecasts

with 23 predictive quantiles. Forecasts from the remaining models are illustrated in Appendix B.7. Forecasts at prediction horizons of three and four weeks are shown in Appendix B.8. All analyses of forecast performance were conducted using the R language for statistical computing (R Core Team, 2021).

3.2.1. Heterogeneity between forecasts

A recurring theme during the evaluation period was pronounced variability between model forecasts. Figure 3.4 illustrates this aspect for point forecasts of incident cases in Germany. The left panel shows the spread of point forecasts issued on 19 October 2020 and valid one to four weeks ahead. The models present very different outlooks, ranging from a return to the lower incidence of previous weeks to exponential growth. The graph also illustrates the difficulty of forecasting cases more than two weeks ahead. Several models had correctly picked up the upwards trend, but presumably a combination of the new testing regime and the semi-lockdown (marked as (a) and (b)) led to a flattening of the curve. The right panel shows forecasts from 9 November 2020, immediately following the aforementioned events. Again, the forecasts are quite heterogeneous. The week ending on Saturday 7 November had seen a slower increase in reported cases than anticipated by almost all models (see Figure 3.2), but there was general uncertainty about the role of saturating testing capacities and evolving testing strategies. Indeed, on 18 November it was argued in a situation report from the Robert Koch Institute (RKI) that the comparability of data from calendar week 46 (9–15 November) to previous weeks is limited (Robert Koch Institute, 2020). This illustrates that confirmed cases can be a moving target and that different modelling decisions can lead to very different forecasts.

Forecasts are not only heterogeneous with respect to their central tendency but also the implied uncertainty. As can be seen from Figures 3.2 and 3.3, certain models issue very confident forecasts with narrow forecast intervals barely visible in the plot. Others – in particular `LANL-GrowthRate` and the exponential smoothing time series model `KIT-time_series_baseline` – show rather large uncertainty. For almost all forecast dates there are pairs of models with no or minimal overlap in 95% prediction intervals, another indicator of limited agreement between forecasts. As can be seen from the right

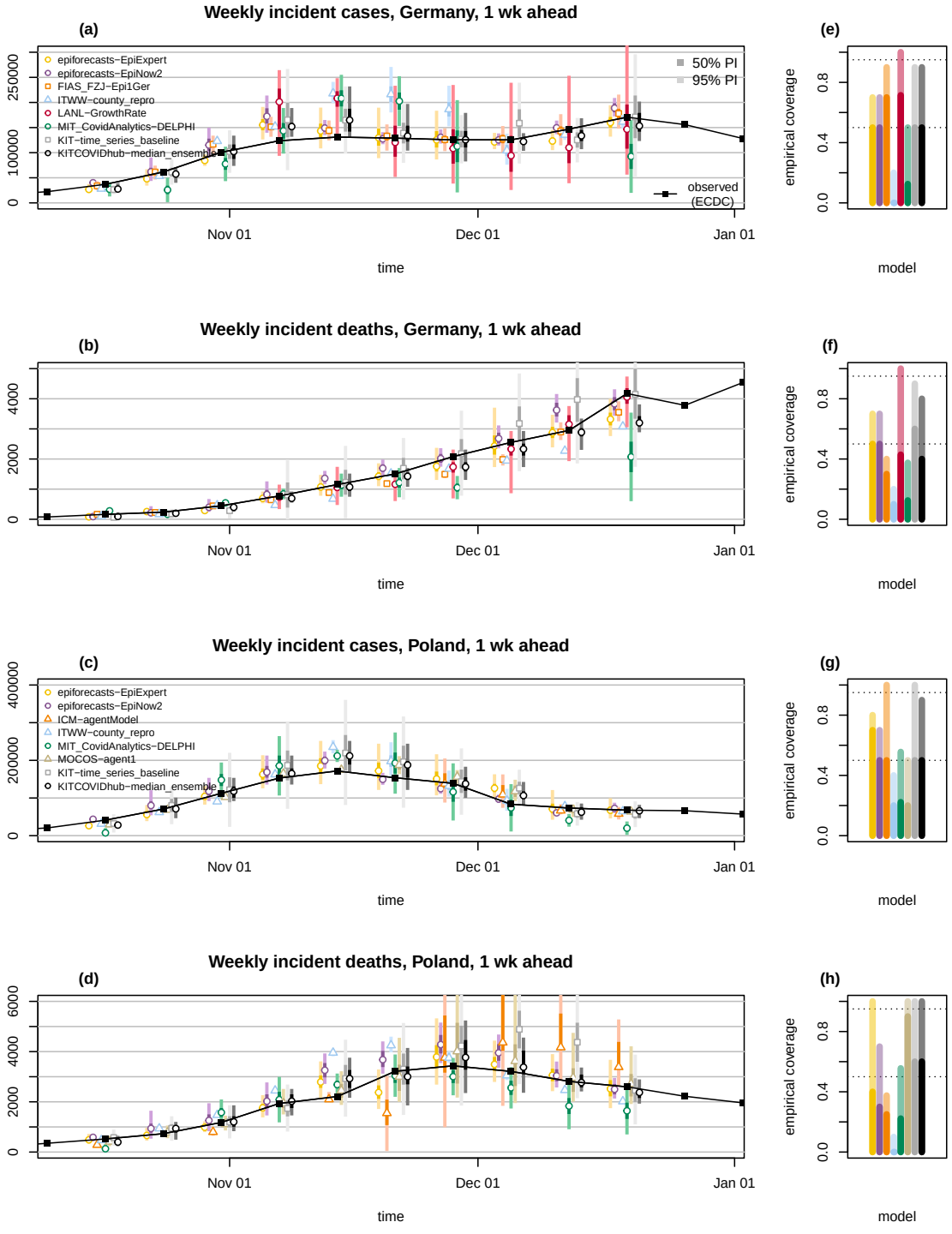


Figure 3.2.: One-week-ahead forecasts. One-week-ahead forecasts of incident cases and deaths in Germany (**a, b**) and Poland (**c, d**). Displayed are predictive medians, 50% and 95% prediction intervals (PIs). Coverage plots (**e**–**h**) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

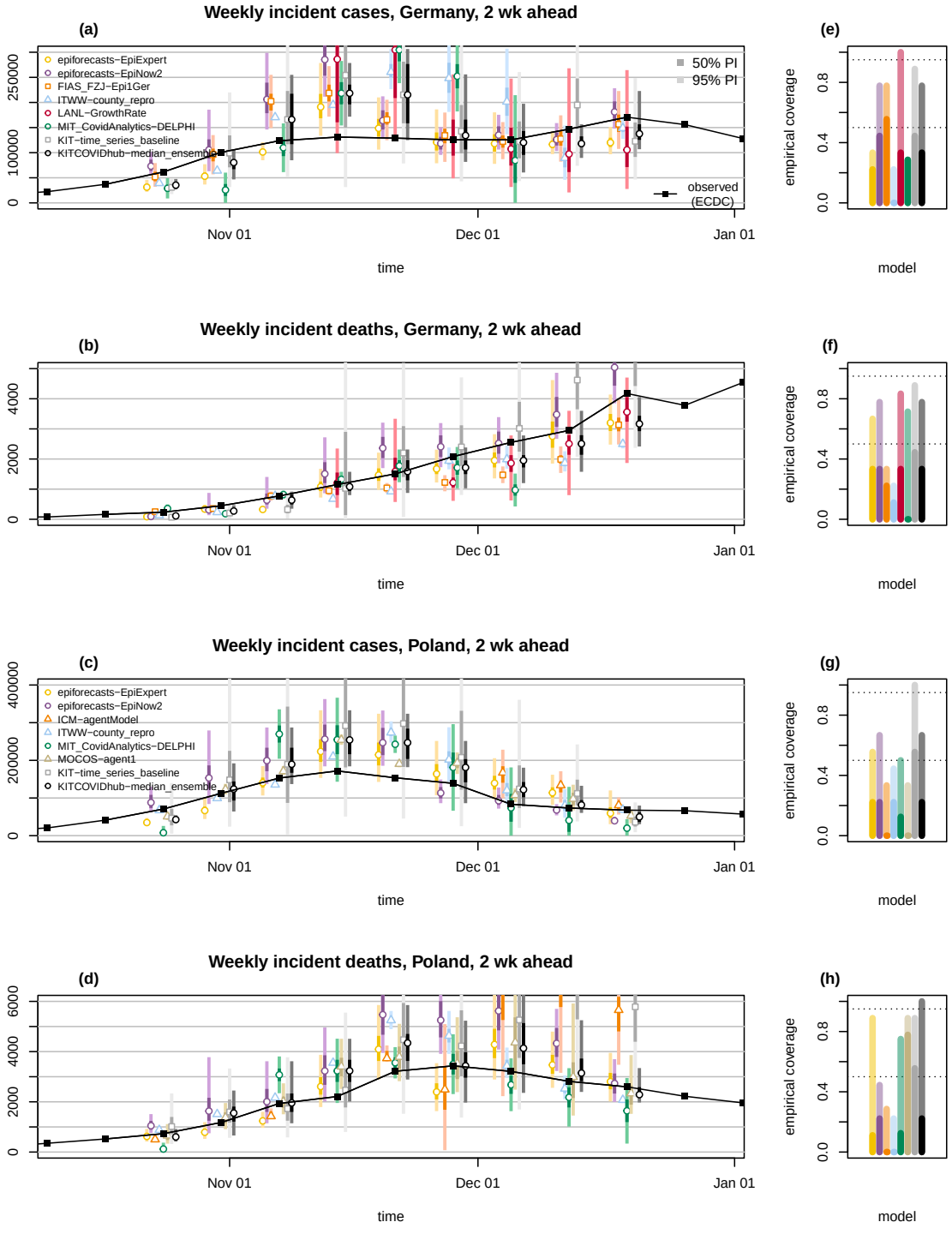


Figure 3.3.: Two-week-ahead forecasts. Two-week-ahead forecasts of incident cases and deaths in Germany (**a, b**) and Poland (**c, d**). Displayed are predictive medians, 50% and 95% prediction intervals (PIs). Coverage plots (**e**–**h**) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

column of Figures 3.2 and 3.3 as well as Tables 3.1 and 3.2, most contributed models were overconfident, i.e. their prediction intervals did not reach nominal coverage.

3.2.2. Adaptation to changing trends and truth data issues

Far from all forecast models explicitly account for interventions and testing strategies (Table 3.3). Many forecasters instead prefer to let their models pick up trends from the data once they become apparent. This can lead to delayed adaptation to changes and explains why numerous models – including the ensemble – showed overshoot in the first half of November when cases started to plateau in Germany (visible from Figure 3.2 and even more pronounced in Figure 3.3). Interestingly, some models adapted more quickly to the flatter curve. This includes the human judgement approach `EpiExpert`, which, due to its reliance on human input, can take information on interventions into account before they become apparent in epidemiological data, but interestingly also `Epi1Ger` and `EpiNow2` which do not account for interventions. In Poland, overshoot could be observed following the peak week in cases (ending on 15 November), with the one-week-ahead median ensemble only barely covering the next observed value. However, most models adapted quickly and were back on track in the following week.

A noteworthy difficulty for death forecasts in Germany was under-prediction in consecutive weeks in late November and December. In November, several models predicted that death numbers would level off, likely as a consequence of the plateau in case numbers starting several weeks before. In the last week of our study (ending on 19 December), most models considerably underestimated the increase in weekly deaths. A difficulty may have been that despite the overall plateau observed until early December, cases continued to increase in the oldest age groups, for which the mortality risk is highest (Figure B.1 in the Appendix). Models that ignore the age structure of cases – which includes most available models (Table 3.3) – may then have been led astray.

A major question in epidemic modelling is how closely surveillance data reflect the underlying dynamics. Like in Germany, testing criteria were repeatedly adapted in Poland. In early September they were tightened, requiring the simultaneous presence of four symptoms for the administration of a test. This was changed to less restrictive criteria in late October (presence of a single symptom). These changes limit the comparability

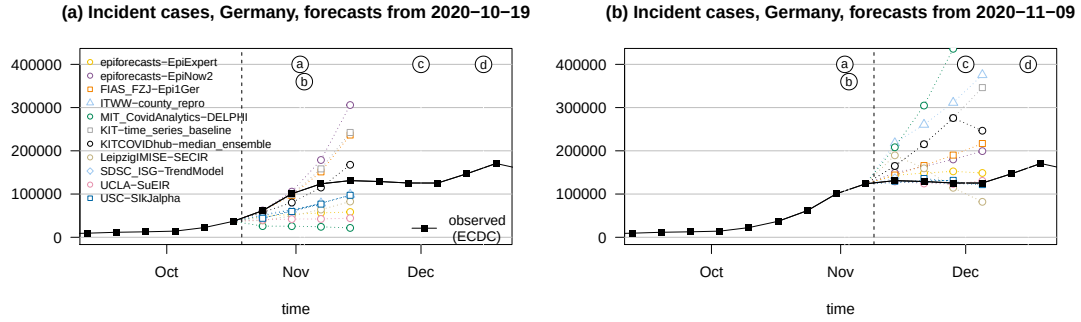


Figure 3.4.: Illustration of heterogeneity between incident case forecasts in Germany. **a** Point forecasts issued by different models and the median ensemble on 19 October 2020. **b** Point forecasts issued on 9 November 2020. The dashed vertical line indicates the date at which forecasts were issued. Events marked by letters a – d are explained in Figure 3.1.

of numbers across time. Very high test positivity rates in Poland (Figure B.2 in the Appendix) suggest that there was substantial under-ascertainment, which is assumed to have aggravated over time. Comparisons between overall excess mortality and reported COVID deaths suggest that there is also relevant under-ascertainment of deaths, again likely changing over time (Afelt et al., 2020). These aspects make predictions challenging, and limitations of ground truth data sources are inherited by the forecasts which refer to them. A striking example of this was the belated addition of 22,000 cases from previous weeks to the Polish record on 24 November 2020. The Poland-based teams `MOCOS` and `MIMUW` explicitly took this shift into account while other teams did not.

3.2.3. Findings for median, mean, and inverse-WIS ensembles

We assessed the performance of forecast ensembles based on various aggregation rules, more specifically a median, a mean, and an inverse-WIS (weighted interval score) ensemble; see the Methods section for the respective definitions.

A key advantage of the median ensemble is that it is more robust to single extreme forecasts than the mean ensemble. As an example of the behaviour when one forecast differs considerably from the others we show forecasts of incident deaths in Poland from 30 November 2020 in Figure 3.5. The first panel shows the six member forecasts, and the

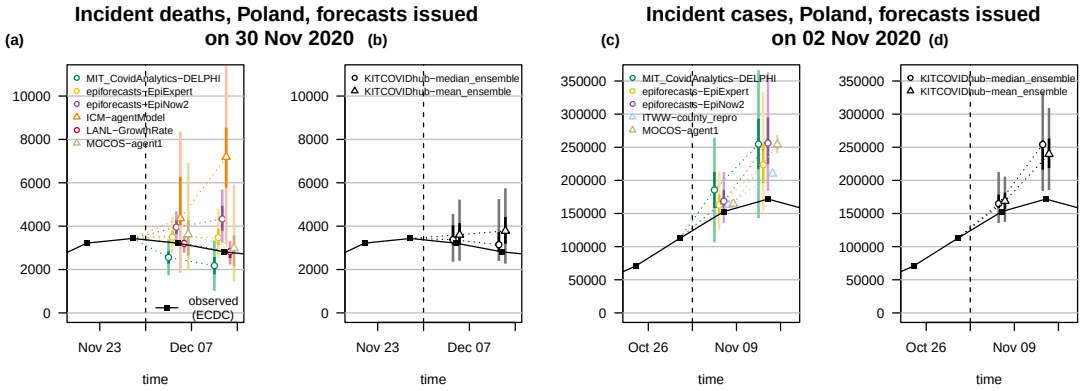


Figure 3.5.: Examples of median and mean ensembles. One- and two-week-ahead forecasts of incident deaths in Poland issued on 30 November, and of incident cases in Poland issued on 2 November 2020. Panels **a** and **c** show the respective member forecasts, panels **b** and **d** the resulting ensembles. Both predictive medians and 95% (light) and 50% (dark) prediction intervals are shown. The dashed vertical line indicates the date at which the forecasts were issued.

second the resulting median and mean ensembles. The predictive median of the latter is noticeably higher as it is more strongly impacted by one model that predicted a resurge in deaths.

A downside of the median ensemble is that its forecasts are not always well-shaped, in particular when a small to medium number of heterogeneous member forecasts is combined. A pronounced example is shown in the third and fourth panel of Figure 3.5. For the one-week-ahead forecast of incident cases in Poland from 2 November 2020, the predictive 25% quantile and median were almost identical. For the two-week-ahead median ensemble forecast, the 50% and 75% quantile were almost identical. Both distributions are thus rather oddly shaped, with a quarter of the probability mass concentrated in a short interval. The mean ensemble, on the other hand, produces a more symmetric and thus more realistic representation of the associated uncertainty.

We briefly address the inverse-WIS ensemble that is a pragmatic approach to giving more weight to forecasts with good recent performance. Figure 3.6 shows the weights of the various member models for incident deaths in Germany and Poland. Note that some models were not included in the ensemble in certain weeks, either because of delayed or missing submissions or due to concerns about their plausibility. While certain models

on average receive larger weights than others, weights change considerably over time. These fluctuations make it challenging to improve ensemble forecasts by taking past performance into account, and indeed Tables 3.1 and 3.2 do not indicate any systematic benefits from inverse-WIS weighting. A possible reason is that models get updated continuously by their maintainers, including major revisions of methodology.

3.2.4. Formal forecast evaluation

Forecasts were evaluated using the mean weighted interval score (WIS), mean absolute error (AE), and interval coverage rates. The WIS is a generalization of the absolute error to probabilistic forecasts and is negatively oriented, meaning that smaller values are better (see the Methods section). Tables 3.1 and 3.2 provide a detailed overview of results by country, target, and forecast horizon, based on data from the European Centre for Disease Prevention and Control (2020b, ECDC). We repeated all evaluations using data from the Center for Systems Science and Engineering at Johns Hopkins University (Dong et al., 2020, JHU CSSE) as ground truth (Appendix B.7), and the overall results seem robust to this choice. We also report on three- and four-week-ahead forecasts in Appendix B.8, though for reasons discussed in the Methods section, we consider their usability limited. To put the results of the submitted and ensemble forecasts into perspective we created forecasts from three baseline methods of varying complexity, see Methods section.

Figure 3.7 depicts the mean WIS achieved by the different models on the incidence scale. For models providing only point forecasts, the mean AE is shown, which as detailed in the Methods section, can be compared to mean WIS values. A simple model always predicting the same number of new cases/deaths as in the past week (`KIT-baseline`) serves as a reference. For deaths, the ensemble forecasts and several submitted models outperform this baseline up to three or even four weeks ahead. Deaths are a more strongly lagged indicator, which favours predictability at somewhat longer horizons. Another aspect may be that at least in Germany, death numbers have been following a rather uniform upward trend over the study period, making it relatively easy to beat the baseline model. For cases, which are a more immediate measure, almost none of the compared approaches meaningfully outperformed the naïve baseline beyond a horizon of one or two weeks. Especially in Germany this result is largely due to the aforementioned overshoot

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

In exploratory analyses (Figure B.9 in the Appendix) we did not find any clear indication that certain modelling strategies (defined via the five categories used in Table 3.3) performed better than others. Following changes in trends, the human judgement model `epiforecasts-EpiExpert` showed good average performance, while growth rate approaches had a stronger tendency to overshoot (Figures B.5–B.8 in the Appendix). Otherwise, the variability of performance within model categories was pronounced and no apparent patterns emerged.

The median, mean and inverse-WIS ensembles showed overall good, but not outstanding relative performance in terms of mean WIS. At a one week lead time, the median ensemble outperformed the baseline forecasts quite consistently for all considered targets, showing less variable performance than most member models (Figures B.5–B.8 in the Appendix). Differences between the ensemble approaches are minor and do not indicate a clear ordering. We re-ran the ensembles retrospectively using all available forecasts, i.e. including those submitted late or excluded due to implausibilities. As can be seen from Table B.4 in the Appendix, this led only to minor changes in performance. Unlike in the US effort (Brooks et al., 2020; Cramer et al., 2022b), the ensemble forecast is not strictly better than the single-model forecasts. Typically, performance is similar to some of the better-performing contributed forecasts, and sometimes the latter have a slight edge (e.g. `FIAS_FZJ-Epi1Ger` for cases in Germany and `MOCOS-agent1` for deaths in Poland). Interestingly, the expert forecast `epiforecasts-EpiExpert` is often among the more successful methods, indicating that an informed human assessment sets a high bar for more formalized model-based approaches. In terms of point forecasts, the

extrapolation approach `SDSC_ISG-TrendModel` shows good relative performance but only covers one-week-ahead forecasts.

The 50% and 95% prediction intervals of most forecasts did not achieve their respective nominal coverage levels (most apparent for cases two weeks ahead). The statistical time series model `KIT-time_series_baseline` features favourably here, though at the expense of wide forecast intervals (Figure 3.2). While its lack of sharpness leads to mediocre overall performance in terms of the WIS, the model seems to have been a helpful addition to the ensemble by counterbalancing the overconfidence of other models. Indeed, coverage of the 95% intervals of the ensemble is above average, despite not reaching nominal levels.

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

Table 3.1.: Detailed summary of forecast evaluation for Germany (based on ECDC data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany, cases

Model	1 wk ahead incident				2 wk ahead incident				1 wk ahead cumulative				2 wk ahead cumulative			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	12,333	8,781	5/10	7/10	30,329	22,157	2/9	3/9	12,334	8,781	5/10	7/10	42,667	30,669	2/9	4/9
epiforecasts-EpiNow2	11,171	7,932	5/10	7/10	37,338	27,293	4/9	7/9	11,171	7,932	5/10	7/10	47,738	34,253	4/9	6/9
FIAS_FZJ-Epi1Ger	7,798	5,709	7/10	9/10	29,190	21,058	5/9	7/9	17,255	11,264	3/10	7/10	38,925	29,937	5/9	6/9
ITWW-county_repro	34,425	28,906	0/10	2/10	64,378	53,136	0/9	2/9	34,077	28,558	0/10	2/10	101,184	84,276	0/9	2/9
LANL-GrowthRate	38,970*	23,379*	5/7	7/7	77,438*	42,294*	2/6	6/6	39,042	26,794	5/10	7/10	116,494	78,224	3/9	5/9
LeipzigIMISE-SECIR	20,019		2/5	3/5	51,115		0/4	1/4	35,901	31,690	1/10	1/10	93,111	83,670	1/9	2/9
MIT_CovidAnalytics-DELPHI	41,313*	29,004*	1/8	4/8	78,872*	61,447*	2/7	2/7								
SDSC_ISG-TrendModel	10,963								10,963							
UCLA-SuEIR	25,012				47,747				25,012				69,800			
USC-SIkJalpha	20,028		1/1	1/1	30,891				21,567		1/1	1/1	49,640			
KIT-baseline	18,475	12,998	5/10	9/10	32,690	25,543	3/9	6/9	18,475	12,998	5/10	9/10	47,472	37,155	3/9	5/9
KIT-extrapolation_baseline	12,016	10,522	7/10	10/10	36,498	26,195	6/9	7/9	12,016	10,522	7/10	10/10	47,145	35,142	7/9	7/9
KIT-time_series_baseline	15,383	11,014	5/10	9/10	44,481	28,625	4/9	8/9	15,383	11,014	5/10	9/10	61,489	39,244	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	14,017	9,358	5/10	9/10	42,063	27,993	2/9	5/9	13,464	9,265	6/10	9/10	52,972	35,194	2/9	8/9
KITCOVIDhub-mean_ensemble	16,649	10,677	4/10	8/10	42,214	27,290	1/9	6/9	15,771	10,630	4/10	8/10	57,125	37,356	1/9	6/9
KITCOVIDhub-median_ensemble	11,534	8,094	5/10	9/10	37,620	25,017	3/9	7/9	12,877	9,243	6/10	7/10	49,438	34,461	2/9	6/9

Germany, deaths

Model	1 wk ahead incident				2 wk ahead incident				1 wk ahead cumulative				2 wk ahead cumulative			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	187	131	5/10	7/10	333	234	3/9	6/9	187	131	5/10	7/10	442	295	3/9	7/9
epiforecasts-EpiNow2	180	120	5/10	7/10	376	235	3/9	7/9	180	120	5/10	7/10	539	336	4/9	7/9
FIAS_FZJ-Epi1Ger	256	223	3/10	4/10	525	433	2/9	3/9	215	175	0/10	4/10	684	564	1/9	4/9
Imperial-ensemble2	254	195	5/10	5/10					252	192	5/10	5/10				
ITWW-county_repro	371	355	1/10	2/10	537	483	1/9	2/9	370	354	1/10	2/10	824	757	2/9	2/9
LANL-GrowthRate	195*	128*	3/7	7/7	457*	313*	2/6	5/6	189	134	3/10	7/10	560	441	3/9	5/9
LeipzigIMISE-SECIR	621		0/5	1/5	768		1/4	1/4	1,167	991	0/10	1/10	2,184	1,799	0/9	1/9
MIT_CovidAnalytics-DELPHI	474*	357*	1/8	3/8	403*	306*	0/7	5/7	449*	331*	1/8	3/8	639*	525*	0/7	3/7
SDSC_ISG-TrendModel	357								357							
UCLA-SuEIR	456				827				456				1,177			
USC-SIkJalpha	489		0/1	0/1	600				500		0/1	0/1	963			
KIT-baseline	479	263	2/10	9/10	835	510	0/9	5/9	479	263	2/10	9/10	1,155	707	0/9	5/9
KIT-extrapolation_baseline	202	134	7/10	9/10	383	246	5/9	8/9	202	134	7/10	9/10	493	330	5/9	8/9
KIT-time_series_baseline	238	190	6/10	9/10	624	415	4/9	8/9	238	190	6/10	9/10	866	577	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	180	114	4/10	9/10	255	147	2/9	8/9	174	108	5/10	9/10	370	224	3/9	8/9
KITCOVIDhub-mean_ensemble	204	138	3/10	9/10	298	174	2/9	8/9	216	147	3/10	9/10	441	262	2/9	8/9
KITCOVIDhub-median_ensemble	200	135	4/10	8/10	334	216	3/9	7/9	202	133	4/10	8/10	440	270	3/9	8/9

*Asterisks mark entries where scores were imputed for at least one week. WIS and AE were imputed with the worst score of 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 3.2.: Detailed summary of forecast evaluation for Poland (based on ECDC data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Poland, cases

Model	1 wk ahead incident				2 wk ahead incident				1 wk ahead cumulative				2 wk ahead cumulative			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	13,643	9,574	7/10	8/10	37,395	24,980	2/9	5/9	13,620	9,596	7/10	8/10	52,523	33,602	2/9	6/9
epiforecasts-EpiNow2	11,006	7,041	5/10	7/10	38,906	25,308	2/9	6/9	11,028	7,048	5/10	7/10	47,373	30,303	2/9	7/9
ICM-agentModel			2/4	4/4			0/3	1/3			1/4	3/4			0/3	2/3
ITWW-county_repro	18,149	14,687	2/10	4/10	33,298	27,208	2/9	4/9	17,227	13,786	3/10	5/10	50,638	41,115	1/9	4/9
LANL-GrowthRate	15,956*	9,490*	3/7	7/7	49,295*	27,220*	1/6	6/6	15,269	9,311	3/10	10/10	62,801	36,562	2/9	8/9
MIMUW-StochSEIR			3/5	5/5			2/4	2/4			2/5	5/5			1/4	2/4
MIT_CovidAnalytics-DELPHI	32,620*	23,266*	2/9	5/9	60,490*	45,815*	1/8	4/8								
MOCOS-agent1	13,273	9,124	2/10	5/10	31,610	24,976	0/9	3/9	13,273	9,124	2/10	5/10	43,215	32,106	1/9	3/9
SDSC_ISG-TrendModel	7,633								7,656							
USC-SIkJalpha	10,292		0/1	1/1	24,138				13,560		0/1	1/1	35,390			
KIT-baseline	28,164	18,119	5/10	9/10	52,890	35,107	2/9	6/9	28,235	18,154	5/10	9/10	80,765	53,656	2/9	6/9
KIT-extrapolation_baseline	18,311	11,917	6/10	10/10	55,060	34,212	3/9	7/9	18,289	11,912	6/10	10/10	76,607	45,935	3/9	8/9
KIT-time_series_baseline	22,497	14,100	5/10	10/10	60,079	37,980	5/9	9/9	22,475	14,098	5/10	10/10	84,530	52,303	4/9	9/9
KITCOVIDhub-inverse_wis_ensemble	12,768	8,456	4/10	9/10	36,229	24,628	3/9	6/9	11,865	7,733	4/10	9/10	44,477	30,643	2/9	6/9
KITCOVIDhub-mean_ensemble	12,982	8,320	3/10	9/10	36,338	23,598	3/9	6/9	12,051	7,582	5/10	10/10	44,254	29,461	2/9	7/9
KITCOVIDhub-median_ensemble	14,196	8,862	5/10	9/10	39,829	24,620	2/9	6/9	14,033	8,698	5/10	9/10	50,935	30,699	2/9	5/9

Poland, deaths

Model	1 wk ahead incident				2 wk ahead incident				1 wk ahead cumulative				2 wk ahead cumulative			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	285	176	4/10	10/10	605	374	1/9	8/9	285	176	4/10	10/10	874	530	3/9	8/9
epiforecasts-EpiNow2	386	261	3/10	7/10	1,110	781	2/9	4/9	386	261	3/10	7/10	1,528	1,049	2/9	5/9
ICM-agentModel	752*	672*	2/8	3/8	1,881*	1,237*	0/7	2/7	1,178*	727*	1/8	4/8	2,955*	1,865*	0/7	3/7
Imperial-ensemble2	397	238	3/10	7/10					369	211	3/10	8/10				
ITWW-county_repro	525	484	0/10	1/10	701	613	0/9	2/9	524	483	0/10	1/10	1,219	1,085	0/9	2/9
LANL-GrowthRate	239*	175*	4/7	7/7	404*	251*	3/6	6/6	216	152	5/10	10/10	637	404	3/9	8/9
MIMUW-StochSEIR			1/5	4/5			1/4	2/4			1/5	4/5			0/4	4/4
MIT_CovidAnalytics-DELPHI	512*	329*	2/9	5/9	663*	434*	1/8	6/8	597*	417*	2/9	6/9	1,075*	782*	1/8	3/8
MOCOS-agent1	194	147	9/10	10/10	420	272	7/9	8/9	194	147	9/10	10/10	556	382	7/9	9/9
SDSC_ISG-TrendModel	154								154							
USC-SIkJalpha	206		0/1	1/1	240				256		0/1	0/1	242			
KIT-baseline	437	275	5/10	10/10	834	529	2/9	7/9	437	274	5/10	10/10	1,245	793	2/9	6/9
KIT-extrapolation_baseline	408	286	6/10	8/10	996	702	5/9	7/9	408	286	6/10	8/10	1,423	989	4/9	7/9
KIT-time_series_baseline	546	339	6/10	10/10	1,371	856	5/9	8/9	546	339	6/10	10/10	1,921	1,212	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	220	153	6/10	10/10	488	313	4/9	8/9	242	162	7/10	10/10	702	450	4/9	9/9
KITCOVIDhub-mean_ensemble	252	163	7/10	9/10	585	362	4/9	8/9	265	171	7/10	9/10	815	522	4/9	9/9
KITCOVIDhub-median_ensemble	215	148	6/10	10/10	471	289	2/9	9/9	231	160	6/10	10/10	707	458	4/9	9/9

*Asterisks mark entries where scores were imputed for at least one week. WIS and AE were imputed with the worst score of 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.

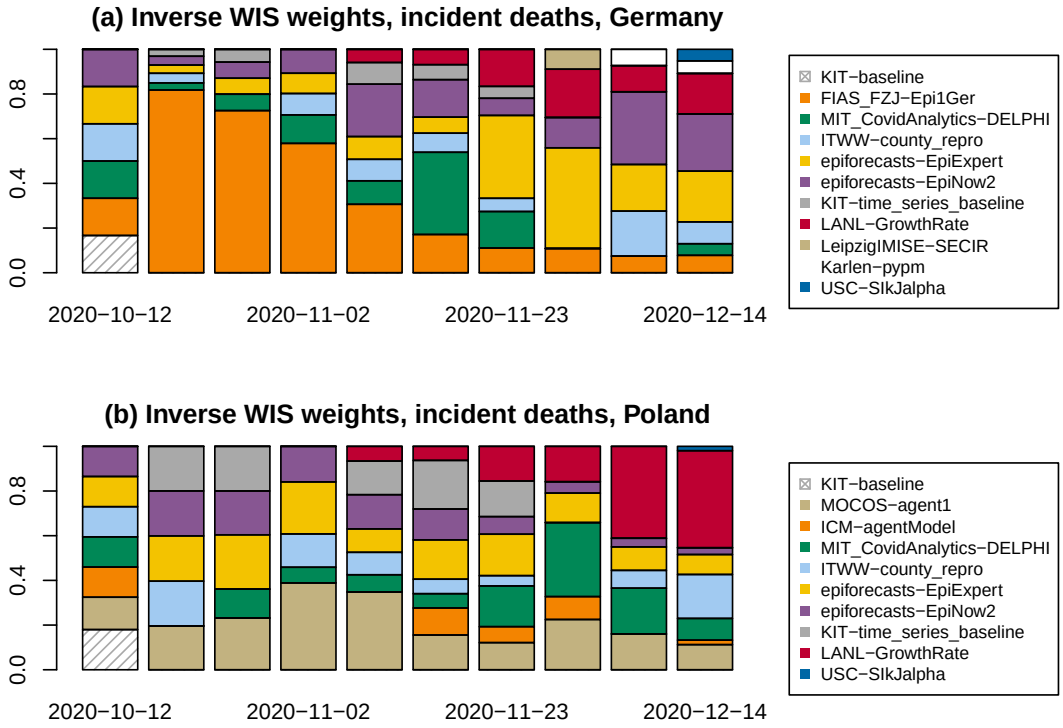


Figure 3.6.: Examples of inverse WIS weights. Inverse-WIS (weighted interval score) weights for forecasts of incident deaths in **a** Germany and **b** Poland.

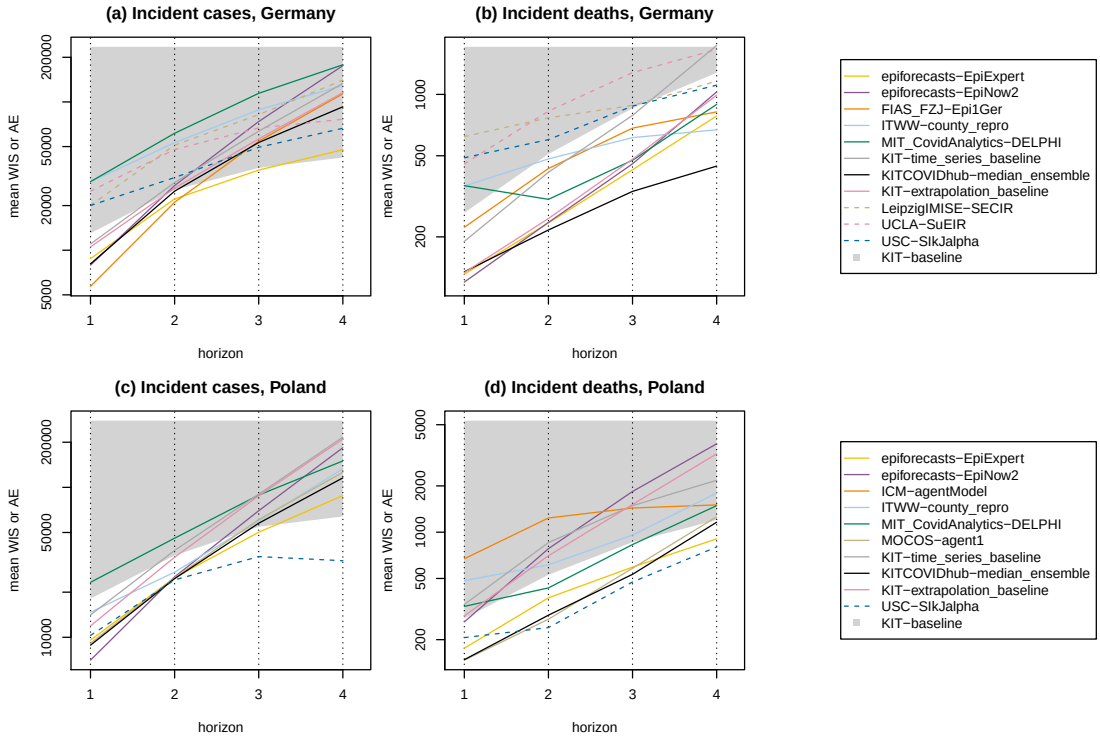


Figure 3.7.: Forecast performance one through four weeks ahead. Mean weighted interval score (WIS) by target and prediction horizon in Germany (**a, b**) and Poland (**c, d**). We display submitted models and the preregistered median ensemble (logarithmic y-axis). For models providing only point forecasts, the mean absolute error (AE) is shown (dashed lines). The lower boundary of the grey area represents the baseline model `KIT-baseline`. Line segments within the grey area thus indicate that a model fails to outperform the baseline. The numbers underlying this figure can be found in Tables 3.1 and 3.2.

3.3. Discussion

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

Similarly to Funk et al. (2020), we conclude that achieving good predictive accuracy and calibration is challenging in a dynamic epidemic situation. Epidemic forecasting is complicated by numerous challenges absent in e.g. weather forecasting (Moran et al., 2016). Noisy and delayed data are an obstacle, but the more fundamental difficulty lies in the complex social (and political) dynamics shaping an epidemic (Funk et al., 2009). These are more relevant for major outbreaks of emerging diseases than for seasonal diseases and limit predictability to rather short time horizons.

Not all included models were designed for the sole purpose of short-term forecasting and could be tailored more specifically to this task. Certain models were originally conceived for what-if projections and retrospective assessments of longer-term dynamics and interventions. This focus on a global fit may limit their flexibility to align closely with the most recent data, making them less successful at short forecast horizons compared to simpler extrapolation approaches. We observed pronounced heterogeneity between the different forecasts, with a general tendency to overconfident forecasting. While over the course of ten weeks, some models achieved better average scores than others, relative performance has been fluctuating considerably.

Various works on multi-model disease forecasting discuss performance differences between modelling approaches, most commonly between mechanistic and statistical approaches. Reich et al. (2019a), McGowan et al. (2019) (both seasonal influenza), and Johansson et al. (2019) (dengue) find slightly better performance of statistical than mechanistic models. All these papers find ensemble approaches to perform best.

Forecasting of seasonal and emerging diseases, however, differ in important ways, the latter typically being subject to more variation in reporting procedures and interventions. This, along with the limited amount of historical data, may benefit mechanistic models. In our study, we did not find any striking patterns, but this may be due to the relatively short study period. We expect that forecast performance is also shaped by numerous other factors, including methods used for model calibration, the thoroughness of manual tuning, and input on new intervention measures or population behaviour.

Different models may be particularly suitable for different phases of an epidemic (Funk et al., 2020), which is exemplified by the fact that some models were quicker to adjust to the slowing growth of cases in Germany. In particular, we noticed that forecasts based on human assessment performed favorably immediately after changes in trends. These aspects highlight the importance of considering several independently run models rather than focusing attention on a single one, as is sometimes the case in public discussions. Here, collaborative forecasting projects can provide valuable insights and facilitate the communication of results. Overall, ensemble methods showed good, but not outstanding relative performance, notably with clearly above-average coverage rates and more stable performance over time. An important question is whether ensemble forecasts could be improved by sensible weighting of members or post-processing steps. Given the limited amount of available forecast history and rapid changes in the epidemic situation, this is a challenging encounter, and indeed we did not find benefits in the inverse-WIS approach.

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

In its present form, the platform covers only forecasts of confirmed cases and deaths. These commonly addressed forecasting targets were already covered by a critical mass of teams when the project was started. Given the limited available time resources of teams, a choice was made to focus efforts on this narrow set of targets. This was also motivated

by the strong focus German legislators have put on seven-day incidences, which have been the main criteria for the strengthening or alleviation of control measures. However, there is an ongoing debate on the usefulness of this indicator, with frequent claims to replace it with hospital admissions (Küchenhoff et al., 2021). An extension to this target was considered, but in view of emerging parallel efforts and open questions on data availability not prioritized. Given that in a post-vaccination setting the link between case counts and healthcare burden is expected to change, however, this decision will need to be re-assessed.

Estimation of total numbers of infected (including unreported) and effective reproductive numbers are other areas where a multi-model approach can be helpful (see UK Department of Health and Social Care (2021) for an example of the latter). While due to the lack of appropriate truth data, these do not qualify as true prediction tasks, ensemble averages can again give a better picture of the associated uncertainty.

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

3.4. Methods

We now lay out the formal framework of our evaluation study. Unless stated differently, the described approach is the same as in the study protocol (Bracher et al., 2020).

3.4.1. Submission system and rhythm

All submissions were collected in a standardized format in a public repository to which teams could submit (`https://github.com/KITmetricslab/covid19-forecast-hub-de`, German and Polish COVID-19 Forecast Hub Team (2021b)). For teams running their own repositories, the Forecast Hub Team put in place software scripts to re-format forecasts and transfer them into the Hub repository. Participating teams were asked to update their forecasts on a weekly basis using data up to Monday. Submission was

possible until Tuesday at 3 pm Berlin/Warsaw time. Delayed submission of forecasts was possible until Wednesday, with exceptional further extensions possible in case of technical issues. Delays of submissions were documented (Table B.1 in the Appendix).

3.4.2. Forecast targets and format

We focus on short-term forecasting of confirmed cases and deaths from COVID-19 in Germany and Poland one and two weeks ahead. Here, weeks refer to Morbidity and Mortality Weekly Report (MMWR) weeks which start on Sunday and end on Saturday, meaning that one-week-ahead forecasts were actually five days ahead, two-week ahead forecasts were twelve days ahead, etc. All targets were defined by the date of reporting to the national authorities. This means that modellers have to take reporting delays into account, but has the advantage that data points are usually not revised over the following days and weeks. From a public health perspective, there may be advantages in using data by symptom onset; however, for Germany, the symptom onset date is only available for a subset of all cases (50–70%), while for Poland no such data were publicly available during our study period. All targets were addressed both on cumulative and weekly incident scales. Forecasts could refer to both data from the European Centre for Disease Prevention and Control (2020b, ECDC) and Johns Hopkins University Center for Systems Science and Engineering (Dong et al., 2020, JHU CSSE). In this article, we focus on the preregistered period of 12 October 2020 to 19 December 2020 (see Figure 3.1). Note that on 14 December 2020, the ECDC data set on COVID-19 cases and deaths in daily resolution was discontinued. For the last weekly data point, we therefore used data streams from the Robert Koch Institute and the Polish Ministry of Health which we had previously used to obtain regional data and which up to this time had been in agreement with the ECDC data.

Most forecasters also produced and submitted three- and four-week-ahead forecasts (which were specified as targets in the study protocol). These horizons, also used in the US COVID-19 Forecast Hub (Ray et al., 2020), were originally defined for deaths. Due to their lagged nature, these were considered predictable independently of future policy or behavioural changes up to four weeks ahead; see UK Scientific Pandemic Influenza Group on Modelling (2020) for a similar argument. During the summer months,

when incidence was low and intervention measures largely constant, the same horizons were introduced for cases. As the epidemic situation and intervention measures became more dynamic in autumn, it became clear that case forecasts further than two weeks (twelve days) ahead were too dependent on yet unknown interventions and the consequent changes in transmission rates. It was therefore decided to restrict the default view in the online dashboard to one- and two-week-ahead forecasts only. At the same time we continued to collect three- and four-week-ahead outputs. Most models (with the exception of `epiforecasts-EpiExpert`, `COVIDAnalytics-Delphi` and in some exceptional cases `MOCOS-agent1`) do not anticipate policy changes, so that their outputs can be seen as "baseline projections", i.e. projections for a scenario with constant interventions. In accordance with the study protocol, we also report on three- and four-week-ahead predictions, but these results have been deferred to Appendix B.8.

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

3.4.3. Evaluation measures

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

$$\mathrm{IS}_\alpha(F, y) = (u - l) \; + \; \frac{2}{\alpha} \times (l - y) \times \chi(y < l) \; + \; \frac{2}{\alpha} \times (y - u) \times \chi(y > u). \qquad (3.1)$$

Here u and l are the lower and upper ends of the respective interval, χ is the indicator function and y is the eventually observed value. The three summands can be interpreted as a measure of sharpness and penalties for under- and overprediction, respectively. The primary evaluation measure used in this study is the weighted interval score (Bracher et al., 2021a, WIS), which combines the absolute error (AE) of the predictive median m and the interval scores achieved for the eleven nominal levels. The WIS is a well-known

quantile-based approximation of the continuous ranked probability score (Gneiting and Raftery, 2007, CRPS) and, in the case of our 11 intervals, defined as

$$\text{WIS}(F, y) = \frac{1}{11.5} \times \left(\frac{1}{2} \times |y - m| \; + \; \sum_{k=1}^{11} \left(\frac{\alpha_k}{2} \times \text{IS}_{\alpha_k}(F, y) \right) \right), \tag{3.2}$$

where $\alpha_1 = 0.02, \alpha_2 = 0.05, \alpha_3 = 0.10, \alpha_4 = 0.20, \ldots, \alpha_{11} = 0.90$. Both the IS and WIS are proper scoring rules (Gneiting and Raftery, 2007), meaning that they encourage honest reporting of forecasts. The WIS is a generalization of the absolute error to probabilistic forecasts. It reflects the distance between the predictive distribution F and the eventually observed outcome y on the natural scale of the data, meaning that smaller values are better. As secondary measures of forecast performance, we considered the absolute error (AE) of point forecasts and the empirical coverage of 50% and 95% prediction intervals. In this context, we note that WIS and AE are equivalent for deterministic forecasts (i.e. forecasts concentrating all probability mass on a single value). This enables a principled comparison between probabilistic and deterministic forecasts, both of which appear in the present study. Applying the absolute error implies that forecasters should report predictive medians, as pointed out in the paper describing the employed evaluation framework (Bracher et al., 2021a).

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

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

3.4.4. Baseline forecasts

In order to put evaluation results into perspective we use three simple reference models. Note that only the first was prespecified. The two others were added later as the need for comparisons to simple, but not completely naïve, approaches was recognized. More detailed descriptions are provided in Appendix B.2.

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

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

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

3.4.5. Contributed forecasts

During the evaluation period from October to December 2020, we assembled short-term predictions from a total of 14 forecast methods by 13 independent teams of researchers. Eight of these are run by teams collaborating directly with the Hub, based on models these researchers were either already running or set up specifically for the purpose of short-term forecasting. The remaining short-term forecasts were made available via dedicated online dashboards by their respective authors, often along with forecasts for other countries. With their permission, the Forecast Hub team assembled and integrated these forecasts. Table 3.3 provides an overview of all included models with brief descriptions and information on the handling of non-pharmaceutical interventions, testing strategies, age strata, and the source used for truth data. More detailed verbal descriptions can be found in Appendix B.3. The models span a wide range of approaches, from computationally expensive agent-based simulations to human judgement forecasts.

Not all models addressed all targets and forecast horizons suggested in our project; which targets were addressed by which models can be seen from Tables 3.1 and 3.2.

3.4.6. Ensemble forecasts

We assess the performance of three different forecast aggregation approaches:

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

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

`KITCOVIDhub-inverse_wis_ensemble` The α-quantile of the ensemble forecast is a weighted average of the α-quantiles of the member forecasts. The weights are chosen inversely to the mean WIS value obtained by the member models over six recently evaluated forecasts (last three one-week-ahead, last two two-week-ahead, last three-week-ahead; missing scores are again imputed by the worst score achieved by any model for the respective target). This is done separately for incident and cumulative forecasts. The inverse-WIS ensemble is a pragmatic strategy to base weights on past performance which is feasible with a limited amount of historical forecast/observation pairs (see Zamo et al. (2021) for a similar approach).

Only models providing complete probabilistic forecasts with 23 quantiles for all four forecast horizons were included in the ensemble for a given target. It was not required that forecasts be submitted for both cumulative and incident targets so that ensembles for incident and cumulative cases were not necessarily based on exactly the same set of models. The Forecast Hub Team reserved the right to screen and exclude member models in case of implausibilities. Decisions on inclusion were taken simultaneously for all three ensemble versions and were documented in the Forecast Hub platform (file decisions_and_revisions.txt in the main folder of the repository). The main reasons for the exclusion of forecasts from the ensemble were forecasts in an implausible order of

Table 3.3.: Forecast models contributed by independent external research teams. Abbreviations: NPI: Does the forecast model explicitly account for non-pharmaceutical interventions? Test: Does the model account for changing testing strategies? Age: Is the model age-structured? DE, PL: Are forecasts issued for Germany and Poland, respectively? Truth: Which truth data source does the model use? Pr: Are forecasts probabilistic (23 quantiles)?

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

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

magnitude or forecasts with vanishingly small or excessive uncertainty. As it showed comparable performance to submitted forecasts, the `KIT-time_series_baseline` model was included in the ensemble forecasts in most weeks.

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

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 (`https://doi.org/10.5281/zenodo.4752079`, German and Polish COVID-19 Forecast Hub Team (2021b)). This repository also contains all truth data used for evaluation. Details on how truth data were obtained can be found in Appendix B.4. Forecasts can be visualized interactively at `https://kitmetricslab.github.io/forecasthub/`. Source data to reproduce all figures are provided with this paper.

Code availability

Codes to reproduce figures and tables are available at `https://github.com/KITmetricslab/analyses_de_pl`, with a stable version at `https://doi.org/10.5281/zenodo.5085398` (German and Polish COVID-19 Forecast Hub Team, 2021a). The results presented in this paper have been generated using the release "revision1" of the repository `https://github.com/KITmetricslab/covid19-forecast-hub-de`, see above for the link to the stable Zenodo release.

Appendix B

B.1. Additional time series plots

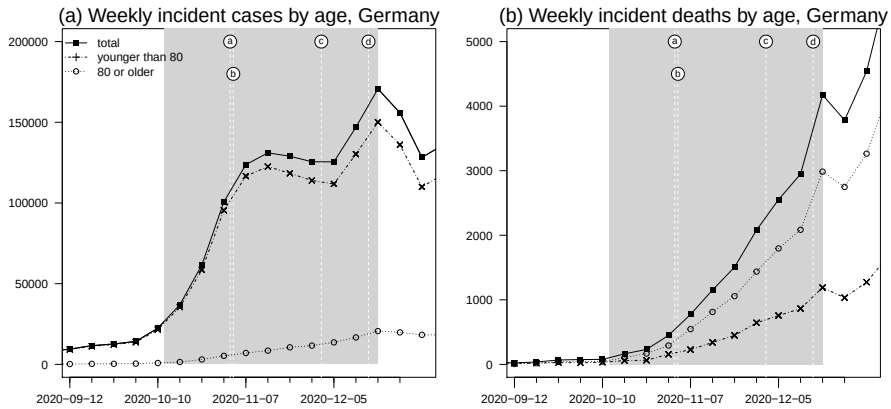


Figure B.1.: **a** Weekly incident COVID-19 cases and **b** deaths in Germany, pooled and stratified by age below and above 80 years. The time period covered by our study is highlighted in grey. Events marked by letters a – d are explained in Figure 3.1.

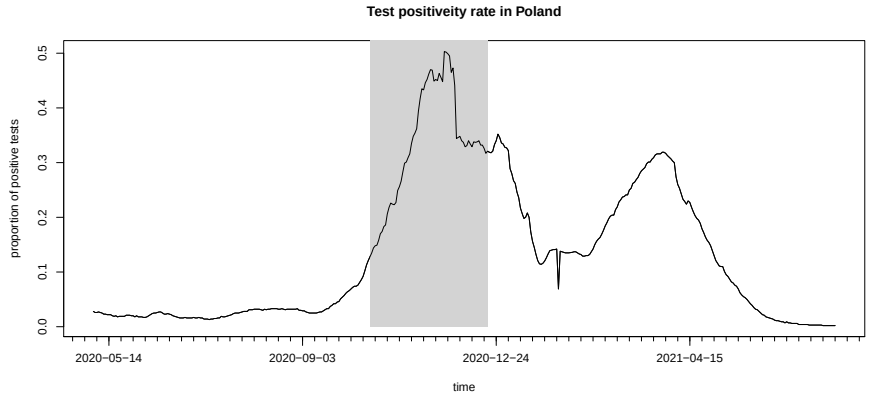


Figure B.2.: Smoothed test positivity rates in Poland. The time period covered by our study is highlighted in grey.

Test positivity data were downloaded from `https://ourworldindata.org/coronavirus-testing` and originate from Hasell et al. (2020).

B.2. Detailed description of baseline forecasts

We here describe the three baseline forecasts mentioned in the main manuscript in more detail.

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

$$\sum_{i=0}^{4} \log \pi(X_{t-i} \mid X_{t-i-1}, \psi) \tag{3.3}$$

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

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

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

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

$$\alpha = \frac{X_t}{X_{t-1}}, \tag{3.4}$$

which corresponds to simple multiplicative extrapolation.

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

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

$$\sum_{i=1}^{5} \log \pi(X_{t-i} \mid \lambda_{t-i}, \psi), \tag{3.5}$$

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

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

`KIT-time_series_baseline` We fit an exponential smoothing model with multiplicative errors and without seasonality to the last 12 observations on the incidence scale. The `R` (R Core Team, 2021) command is

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

using the `forecast` package (Hyndman and Khandakar, 2008) (version 8.12). As noted in the main text, this specification is taken from Petropoulos and Makridakis (2020). As in the previous section we proceed by sampling paths from this model and computing predictive quantiles from them.

B.3. Descriptions of submitted forecasts

We provide more extensive descriptions of the models listed in Table 3.3, including details on inference approaches and computation times.

`CovidAnalytics-DELPHI` Predictions for future cases are obtained from a heavily modified SEIR model. New states are added to account for cases that went undetected, while quarantined and hospitalized patients are separated. The infection rate is corrected with a nonlinear curve that represents the cumulative effect of governmental and societal response (which is assumed to change according to the magnitude of the outbreak). Key parameters for the disease are fixed using a meta-analysis conducted by the CovidAnalytics group of over 150 parameters while epidemiological parameters are fitted to historical death counts and detected cases. The epidemiological parameters are fitted using a moving time window with truncated Newton and simulated annealing to effectively capture the latest changes in the epidemic trends. Uncertainty intervals are generated using re-sampling of the out-of-sample error for the predictions two weeks ago by fitting the historical out-of-sample error to a normal distribution. Then we assume that the incremental estimates for each week have random errors of such magnitude that are independent (and additive) by week. Generating the forecasts for one country takes a few minutes on a standard laptop.

`epiforecasts-EpiExperts` The `EpiExpert` model represents predictions from experts and non-experts that were submitted through an R Shiny application (`app.crowdforecastr.org`). Participants could select a distribution family and specify the median and spread of the predictive distribution by dragging points or adjusting numeric input values. Available distributions were several transformations of a normal distribution (normal, log-normal, 3rd-, 5th, and 7th-power normal). From every predictive distribution, 23 quantiles were obtained. Lastly, an ensemble was formed by aggregating predictions using a quantile-wise mean. Participation increased over time from around 3 to around 5–10 weekly forecasters. To inform their forecasts, participants could use any data they liked. The data shown in the app was data from the ECDC, RKI and the Polish Ministry of Health. Participants were shown additional information from `ourworldindata.org`.

epiforecasts-EpiNow2 EpiNow2 is an exponential growth model that uses a time-varying R_t trajectory to forecast latent infections, and then convolves these infections using known delays to observations, via a negative binomial model coupled with a day of the week effect. It makes limited assumptions and is not tuned to the specifics of COVID-19 in Germany and Poland beyond epidemiological details such as literature estimates of the generation time, incubation period and the population of each country. The reproductive number R_t is assumed to remain static after the respective last observed values.

Each forecast target was fit independently for each model using Markov-chain Monte Carlo (MCMC) using the Stan software and RStan interface (https://mc-stan.org/rstan/). A minimum of 4 chains were used with a warmup of 250 samples and 2000 samples total post warmup. Forecast intervals were calculated from the generated MCMC samples, aggregating daily posterior samples to the required weekly scale. Forecast generation for all targets, both national and subnational, took not more than 2 hours using 16 CPU cores.

FIAS FZJ-Epi1Ger This is an extended version of the well-established SEIR (susceptible – exposed – infectious – removed) model, i.e. a deterministic approach based on a system of ordinary differential equations. Mixing in the population is assumed to be homogeneous, but infectiousness varies across three possible compartments, namely, asymptomatic undetected, symptomatic undetected and detected cases. Detection can occur both in an early stage, that is, during latent phase (E, assumed not yet infectious) or when the patient is already infectious (asymptomatic/symptomatic). Undetected infections lead to undetected recoveries, whereas detected cases may either recover or die. We do not explicitly incorporate a reporting delay in recording fatalities. Daily new detected cases and reported deaths are used for model calibration. Transmission, detection and death rates are assumed to be piecewise constant in time and fitted using reported data (time series data from RKI/ECDC). The fitting algorithm is based on Monte Carlo sampling and minimization of sum of squared residuals, evaluating results with the Akaike information criterion. Parameter values obtained from the fit of the most recent interval are used for model predictions (incident/cumulative cases and deaths). The model is not

age-stratified and does not take into account advance knowledge on mitigation measures to be applied, e.g. to reduce contact rates in the population. To quantify the uncertainty of forecasts, the model fits obtained via Monte Carlo sampling of the parameter space are collected in histograms according to their Akaike weights. Predictive quantiles are calculated from these weighted histograms. The calculations take about 10 to 15 minutes on a single core of a workstation, but can be accelerated considerably by partly re-using results of recent model runs.

`ICM-agentModel` The model aims to represent the social structure of the Polish population at the level of the individual citizens and their social contacts. It is spatially structured and contains representations of different contexts. The model follows the development of the epidemic on a geographical grid with a resolution of 1km and through physical contact in various contexts of social life. It currently contains representations of households, workplaces, kindergardens, schools, universities, travel, streets and public places. The approach can serve for short-term prediction of the development of the epidemic and for the exploration of various scenarios. These can shed light on the effects of newly introduced policies and non-pharmaceutical interventions. The model features a detailed demographic stratification where each agent has an attribute of age and gender. The likelihood of severe symptoms, hospitalization, ICU treatment and death depends on age. Agents of different age appear in different contexts (schools, workplaces, travel, etc.) with different probabilities. The model is fitted via a Bayesian inference approach. Data inputs include publicly available epidemiological data (number of cases, deaths etc) from the Polish Ministry of Health as well as more detailed data available via the Polish National Institute for Public Health (time of hospitalization, age, symptoms, contact tracing etc). Computations are performed on a Cray-XC40 supercomputer. The generation of forecast for Poland takes 1 hour of computing time on 10 nodes (node: Intel Xeon E5-2690 v3 2.6 GHz 2 x 12 cores, hyperthreading x2, 128 GB RAM).

imperial-ensemble2 To generate forecasts at a horizon of seven days, an unweighted ensemble of three models is used:

- Model 1 assumes a conditional Poisson distribution and the renewal equation

$$I_t \sim \text{Pois}\left(R_t \sum_{s=0}^{t} I_{t-s} w_s\right), \tag{3.6}$$

  where I_t are deaths reported on day t. Here, the serial interval distribution w is a gamma distributon with mean 6.48 days and a standard deviation of 3.83 days. The instantaneous reproduction number R_t is estimated via a maximum likelihood approach from the last 10 days (jointly with the incidence prior to the 10 day window). Forecasts are obtained by simulating stochastic realisations of the renewal equation from the end of the calibration period.

- Model 2 optimises the choice of time window over which the reproduction number is assumed to be constant for estimating. The optimal window is one which minimises the accumulated predictive error (APE) in 1-step ahead predictions over the entire time series. Estimates of reproduction numbers using the optimal time window are then used to project forward using the renewal equation.

- Model 3 uses both the reported number of cases and deaths. The reported cases are weighted with a reporting to death delay distribution to obtain the largest potential number of deaths, if all reported cases were to die. The observed number of deaths is used to estimate an observed fatality ratio. Forward projections are then obtained by sampling from a binomial distribution with the weighted case count and estimated fatality ratio.

For each model, we generate 10000 samples from the posterior distribution of reproduction number and obtain 10000 simulations of forward projections. Uncertainty intervals are provided by the quantiles of the posterior distribution. Computations require approximately 10 minutes on an 8 GB Macbook Pro.

ITWW-county_repro Reproduction numbers on the county level are estimated via a small area approach where reproduction numbers are modeled as random variables. First we sample for every county reproduction numbers for the past week and compute their mean. This value is held constant over one simulation and we sample from the reproduction equation to generate daily future cases on the county level. These are aggregated to the state and country level. Finally, for Germany, we account for the delay with which these cases appear in the ECDC dataset as our reproduction numbers are based on the official RKI dataset. To forecast deaths we sample on the state level the age of every case and use case-fatality ratios for every age group to forecast which cases will later be marked as death; the delay until death is sampled from a reporting-to-death distribution. We then aggregate the deaths from the state level to the country level. The model contains a detailed age stratification. For Germany we use age-stratified attack-rates and case-fatality ratios which we estimate from the official RKI data. The age groups we consider are: 0–4, 5–14, 15–34, 35–59, 60–79, 80+ and unknown. For Poland we consider age groups from 0 to 100 by year (101 age groups). Generation of forecasts takes less than 10 minutes on a standard laptop.

LANL-GrowthRate The model makes predictions about the future, unconditional on particular intervention strategies. It consists of two processes. The first process is a statistical model of how the number of COVID-19 infections changes over time. The second process maps the number of infections to the reported data. We model the growth of new cases as the product of a dynamic growth parameter and the underlying numbers of susceptible and infected cases in the population at the previous time step, scaled by the size of the state's starting susceptible population. The growth parameter can be thought of as the transmissibility of the virus in that state on that date and is a weighted regression between the trend in the growth rate over the past 42 days and a growth rate that would keep the number of new daily confirmed cases constant. The weights of these two components are dynamically tuned to the observed data. To model new deaths in the population, we assume that a fraction of the 1, 2, 3, 4, or 5-week moving average of the daily confirmed cases will die. The model learns both the moving average window and the case fatality fraction that best fits the historical observations.

LeipzigIMISE-SECIR An adapted mechanistic epidemiological model of the SECIR (susceptible – exposed – carriers – infected – recovered) type is integrated into Input-Output Non-Linear Dynamical Systems (IO-NLDS) serving as hidden layers, i.e. the true dynamics cannot be observed directly. The model contains an asymptomatic compartment, a compartment of patients requiring intensive care, and most of the compartments are divided into three sub-compartments to represent time delays. Changing factors of the system due to non-pharmaceutical interventions, changing age-structure of the infected population, and changes in testing policies are imposed as inputs to the system. Parameters are then estimated by a knowledge synthesis process considering parameter ranges derived from external studies and public data. Specifically, a Bayesian inference approach is taken to estimate time-varying parameters. Public data is translated to model outputs not identical but related to hidden states of the model. The model is fitted to data by a full information approach. The Stochastic Approximation Expectation Maximization (SAEM) algorithm is used to estimate model parameters by minimizing the negative log-likelihood of the observations. The number of updates of time-dependent variables is determined via the Bayesian information criterion (BIC). After determination of residual errors of parameters via SAEM, MCMC chains are simulated in order to sample possible alternative parameter sets around the optimal solution from the distribution, determined by the constrained negative log-likelihood function. Generation of forecasts requires several hours on a standard desktop computer.

MIMUW-StochSEIR The model uses an extension of the SEIR (susceptible – exposed – infected – recovered) approach. The key developments to account for the specifics of COVID-19 are: (i) inclusion of the daily number of tests into the model; (ii) addition of a state representing undiagnosed infected and undiagnosed recovered patients; (iii) a Bayesian inference approach. The main parameters of this extended model are assigned prior distributions with hyperparameters based on the literature or preliminary analyses using deterministic SEIR models. The posterior distributions are computed using the Monte Carlo Metropolis–Hastings algorithm. The final parameter estimates are then given by the posterior means, and the uncertainty intervals by respective quantiles. The

typical computation (3000 steps with 1000 steps of the burn-in phase) takes roughly 4 hours on a 2.3 GHz Dual-core Intel Core i5.

`MOCOS-agent1` This is an agent based model based on heterogeneous random network structures for potentially infectious contacts. The network structure is defined by sets of context and feature-dependent non-symmetric kernels. The dynamics is time-continuous, event driven microsimulation. It takes into account census data on household composition, age distribution, work places etc. The model includes contact tracing, both classic and app-based, testing and quarantine. All relevant duration times like incubation time, time to hospitalization and time to testing are sampled from distributions based on empirical data. Spatial structures are implemented but not used at the moment for country wide forecasts. We take into account changes in delay distribution over time, e.g in time from symptom onset to reporting. Even more importantly, changes in testing strategies and their impact on the dark figure of cases are taken into account

Inference is done using a mixture of Baysian inference, maximum likelihood methods and Monte Carlo search. Uncertainty intervals are generated via a likelihood distribution over an ensemble of suitable sample paths involving different parameters. For each relevant parameter configuration, 100 sample paths are generated and weighted.

Computations are done in parallel on the Lower Silesia (Poland) high performance super cluster and depending on size of parameter space to be searched can take more than 24 hours

`UCLA-SuEIR` The SuEIR model is a variation of the SEIR (susceptible – exposed – infected – recovered) model that features the following compartments (with the total size of the population denoted by N):

- S_t: The number of susceptible individuals at time t, i.e. individuals who can still acquire the disease.
- E_t: The number of individuals who have already been infected but have not been tested/diagnosed with the disease. Unlike in the classic SEIR model this group can also cause new infections.
- I_t: The number of individuals who are infected and have received a positive test.

- R_t: The number of individuals who have been infected and received a positive test and have since either recovered or died.
- u_t: The unobserved number of individuals who have been infected, but not received a positive test, and have since recovered or died.

These compartments are linked via the following system of differential equations:

$$\frac{\mathrm{d}S_t}{\mathrm{d}t} = -\frac{\beta(I_t + E_t)S_t}{N}, \quad \frac{\mathrm{d}E_t}{\mathrm{d}t} = \frac{\beta(I_t + E_t)S_t}{N} - \sigma E_t, \quad \frac{\mathrm{d}I_t}{\mathrm{d}t} = \mu\sigma E_t - \gamma I_t, \tag{3.7}$$

$$\frac{\mathrm{d}R_t}{\mathrm{d}t} = \gamma I_t, \quad \frac{\mathrm{d}u_t}{\mathrm{d}t} = (1-\mu)\sigma E_t, \tag{3.8}$$

where β is the contact rate between the susceptible and infectious (i.e. I_t and E_t) groups, γ is the rate at which individuals leave the detected case compartment (due to recovery or death), μ is the discovery rate and σ is the ratio of cases in the exposed compartments that are either confirmed as infectious or dead/recovered without confirmation. To generate predictions of fatalities, a time-varying ratio r_t of deaths and removals due to recovery is estimated. The model is fit to reported cases and fatalities via a logarithmic-type mean squared error approach and gradient based optimization.

`USC-SIkJalpha` Forecasts are generated using an epidemic model called `SIkJalpha`, a preliminary version of which has been successfully used during the DARPA Grand Challenge 2014. The model can consider the effect of many complexities of the epidemic process and yet be simplified to a few parameters that are learned using fast linear regressions. Therefore, the approach can learn and generate forecasts extremely quickly. The model is able to quickly adapt to changing trends, and the variations in parameters during different times/policies enable the generation of different scenarios such as what would happen if we were to disregard social distancing suggestions. For each state, hospitalizations are modelled as a separate compartment, as a linear function of recent cases with heterogeneous rates. This means that for a hyper-parameter J, those who were infected at time $t-1$ to $t-J$ will have a separate rate from those infected at $t-(J+1)$ to $t-2J$, and so on. Death forecasts are generated in a similar fashion. For long-term forecasts (more than a few days in the future), cases are predicted first based on the `SIkJalpha` model, which forms the input to hospitalization prediction. While

changing trends are accounted for by putting more emphasis on recently seen data, it is assumed that the trends will remain the same in the future for point forecasts. The approach attempts to account for changing trends in the future in the quantile forecasts by modeling the empirical errors using a random forest.

`SDSC-ISG_TrendModel` Forecasts are based on the reported numbers of cases and deaths at the country or regional level. No information about measures or changes in policies is used for forecasting. The modelling substantially relies on trend estimation of the time series of the number of daily cases/deaths. To account for non-stationary weekly seasonality, outliers, missing data and delayed reports, and reliably estimate the underlying trend for each of the time series, we use a robust seasonal trend decomposition model based on LOESS (LOcally Estimated Scatterplot Smoothing). In order to better adapt to the changes in the data, trends are estimated locally in overlapping time series subintervals, while the global smooth trend estimate is a pointwise weighted combination of the overlapping local models. To predict daily cases and deaths we use linear extrapolation of the estimated smooth trend either on the original or on the logarithmic scale.

B.4. Details on truth data sources

The different truth data are publicly available in the following locations:

- Daily data compiled by ECDC until 14 December are available at https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide
- Daily data compiled by the Center for Systems Science and Engineering at Johns Hopkins University (Dong et al., 2020, JHU) are available at https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data
- Daily data from Robert Koch Institute were extracted from https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0. Note that the data provided there are on a different time scale than used in this article (a mix of symptom onset and date of reporting to local authorities). Data by reporting date to the national authorities were generated by comparing data sets made available on subsequent days. The generated data set (available in the Forecast Hub platform) is in agreement with the ECDC data up to 14 December.
- Daily data from the Polish Ministry of Health were extracted from a widely used public spread sheet maintained by Michal Rogalski: https://docs.google.com/spreadsheets/u/2/d/1ierEhD6gcq51HAm433knjnVwey4ZE5DCnu1bW7PRG3E. These data are in agreement with the ECDC data up to 14 December.

All truth data time series, including historical versions, are also available in the folder https://github.com/KITmetricslab/covid19-forecast-hub-de/tree/master/data-truth of our repository.

B.5. Sources on changes in non-pharmaceutical interventions and testing regimes

We here provide sources for the dates of interventions shown in Figure 3.1.

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

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

Specific news items on mentioned interventions/events:

- Four symptoms required for test: Ministerstwo Zdrowia przekazało zasady zlecania testów na koronawirusa, Wprost, 23 September 2020, https://www.wprost.pl/koronawirus-w-polsce/10368723/ministerstwo-zdrowia-przekazalo-zasady-zlecania-testow-na-koronawirusa.html (last accessed 22 December 2020).
- Only one out of four symptoms required for test: Dlaczego lekarz odmawia skierowania na test na COVID-19? Medonet, 5 November 2020, https://www.medonet.pl/koronawirus/koronawirus-w-polsce,kiedy-lekarz-moze-odmowic-skierowania-na-test-na-koronawirusa,artykul,26303647.html (last accessed 22 December 2020)
- Bulk reporting of 22,000 cases on 24 November: Rozbieżności w statystykach koronawirusa. 22 tys. przypadków będą doliczone do ogólnej liczby wyników, Forsal, 23 November 2020, https://forsal.pl/lifestyle/zdrowie/artykuly/8017628,rozbieznosci-w-statystykach-koronawirusa-22-tys-przypadkow-beda-doliczone-do-ogolnej-liczby-wynikow.html (last accessed 22 December 2020)
- High test positivity and suspected under-ascertainment: Polish doctors fear high rate of positive COVID tests show pandemic worse than it appears, J. Plucinska, Reuters, 1 December 2020, https://www.reuters.com/article/us-health-c

oronavirus-poland-cases/polish-doctors-fear-high-rate-of-positive-covid-tests-show-pandemic-worse-than-it-appears-idUSKBN28B54Q (last accessed 22 December 2020)

Germany: A chronicle of the most important events (in German) can be found on the web site of the German Ministry of Health and in a report issued by Robert Koch Institute:

- https://www.bundesgesundheitsministerium.de/coronavirus/chronik-coronavirus.html
- Schilling, J., Buda, S., Fischer, M., Goerlitz, L., Grote, U., Haas, W., Hamouda, O., Prahm, K., and Tolksdorf, K. (2021). Retrospektive Phaseneinteilung der COVID-19-Pandemie in Deutschland bis Februar 2021. *Epidemiologisches Bulletin*, 2021/15:3–12. Available at http://dx.doi.org/10.25646/8149.

Specific news items on mentioned interventions/events:

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

B.6. Availability and delays of forecasts

Table B.1.: Availability of forecasts by model, target and forecast horizon. Each entry describes up to which forecast horizon (in weeks) forecasts for incident cases, cumulative cases, incident death and cumulative deaths were made available (numbers in this order and separated by semicolons). Asterisks indicate that forecasts were only available on Wednesday or later rather than before Tuesday 3pm.

Germany

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

Poland

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

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

B.7. Additional results for one- and two-week-ahead forecasts

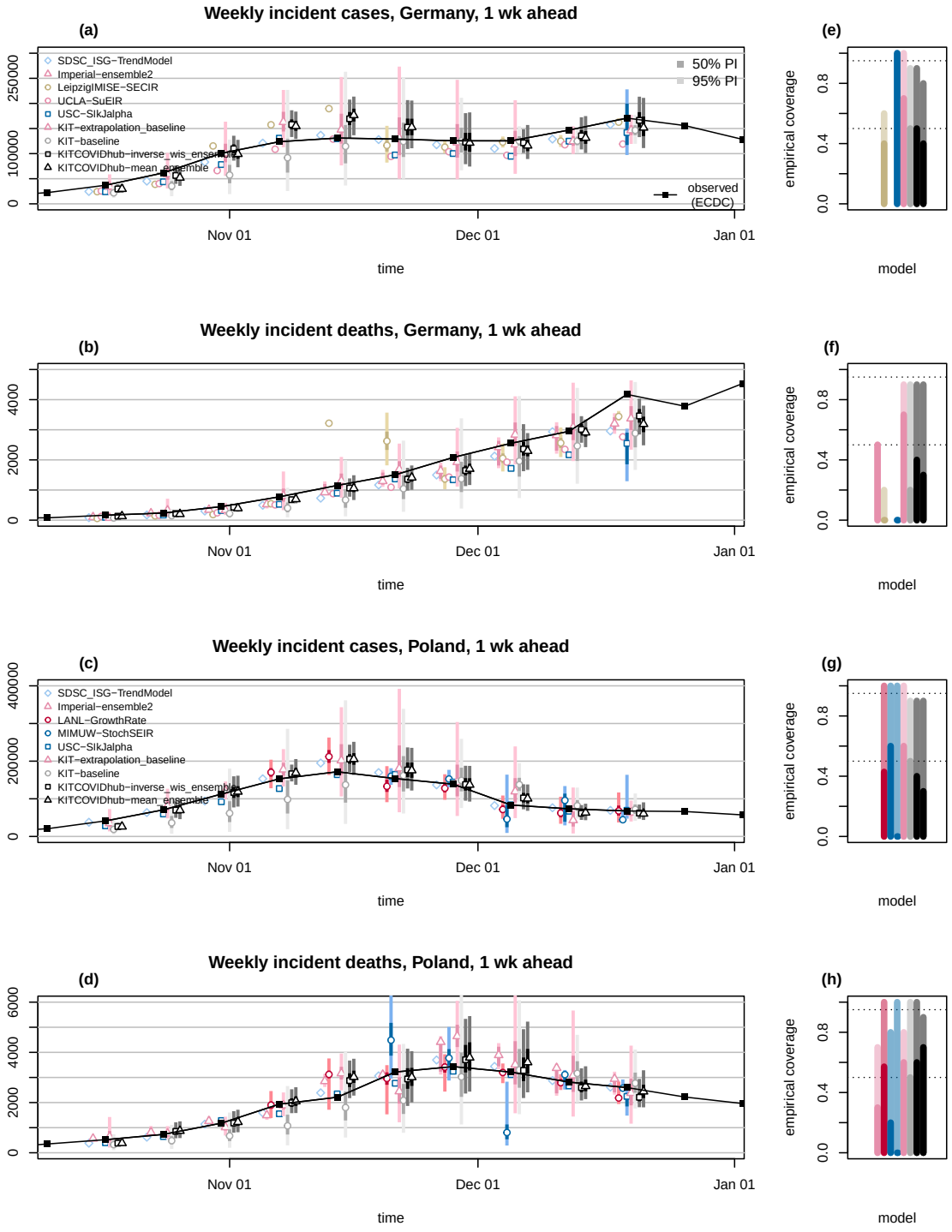


Figure B.3.: Additional one-week-ahead forecasts. One-week-ahead forecasts of incident cases and deaths in Germany (**a, b**) and Poland (**c, d**). Displayed are predictive medians, 50% and 95% prediction intervals (PIs) for models not shown in Figure 3.2. Coverage plots (**e**–**h**) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

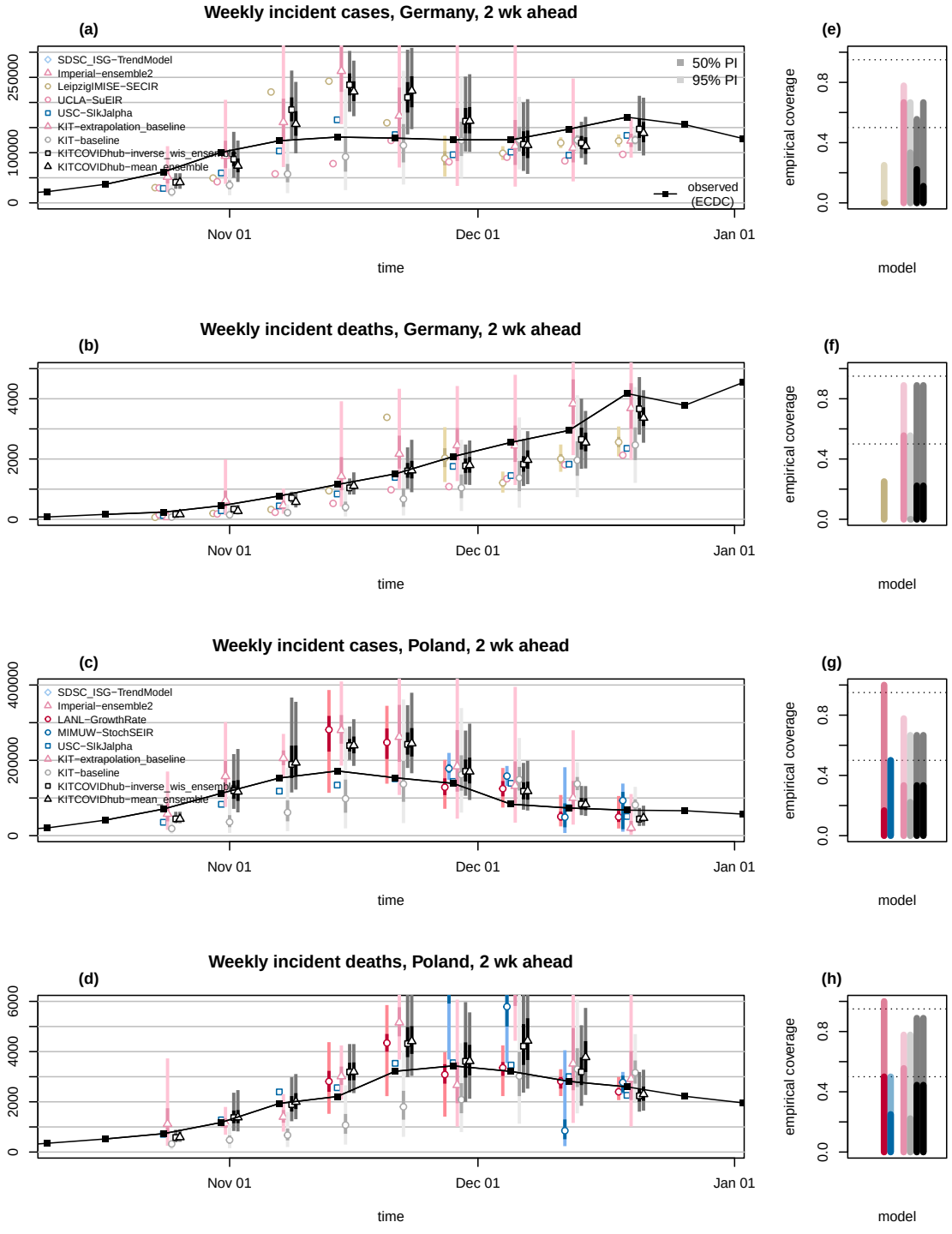


Figure B.4.: Additional two-week-ahead forecasts. Two-week-ahead forecasts of incident cases and deaths in Germany (**a, b**) and Poland (**c, d**). Displayed are predictive medians, 50% and 95% prediction intervals (PIs) for models not shown in Figure 3.3. Coverage plots (**e**–**h**) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

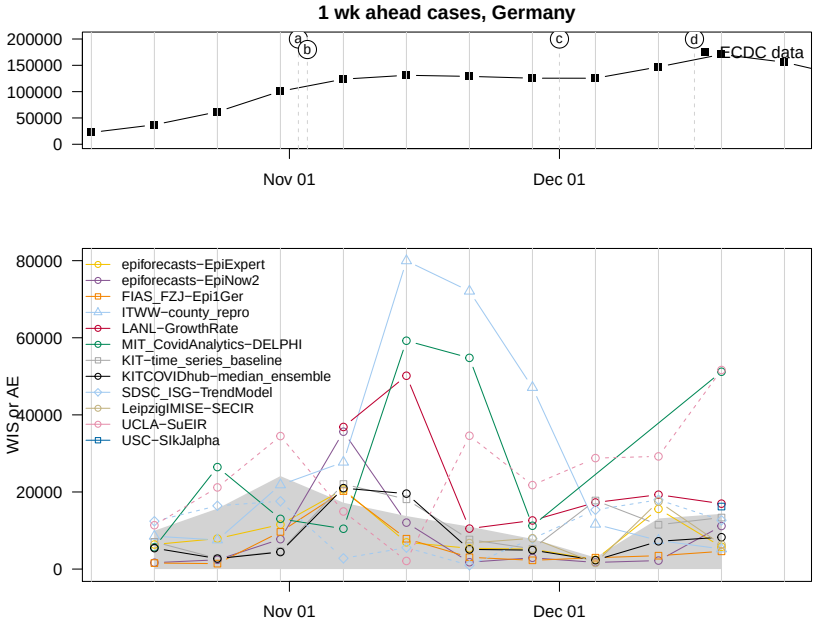


Figure B.5.: WIS value or absolute error (of deterministic models) over time, for one-week-ahead case forecasts for Germany. Letters in circles represent events explained in Figure 3.1. As in Figure 3.7, the upper border of the light grey area represents the performance of the naïve baseline model `KIT-baseline`.

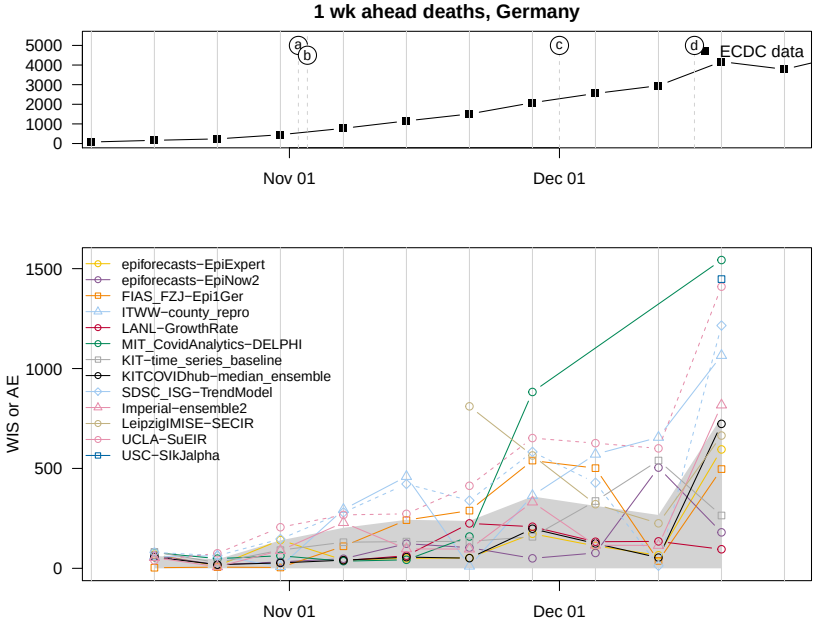


Figure B.6.: WIS value or absolute error (of deterministic models) over time, for one-week-ahead death forecasts for Germany. Letters in circles represent events explained in Figure 3.1. As in Figure 3.7, the upper border of the light grey area represents the performance of the naïve baseline model `KIT-baseline`.

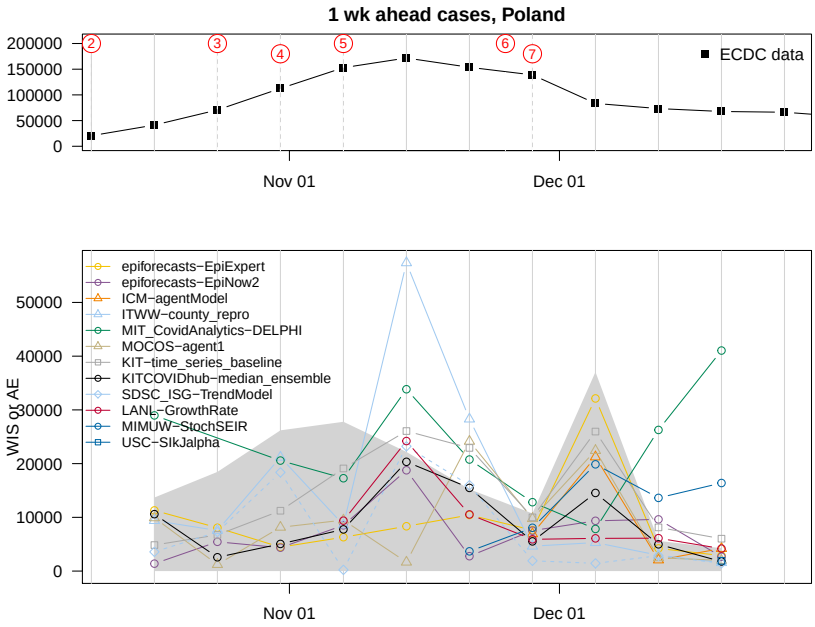


Figure B.7.: WIS value or absolute error (of deterministic models) over time, for one-week-ahead case forecasts for Poland. Numbers in circles represent events explained in Figure 3.1. As in Figure 3.7, the upper border of the light grey area represents the performance of the naïve baseline model `KIT-baseline`.

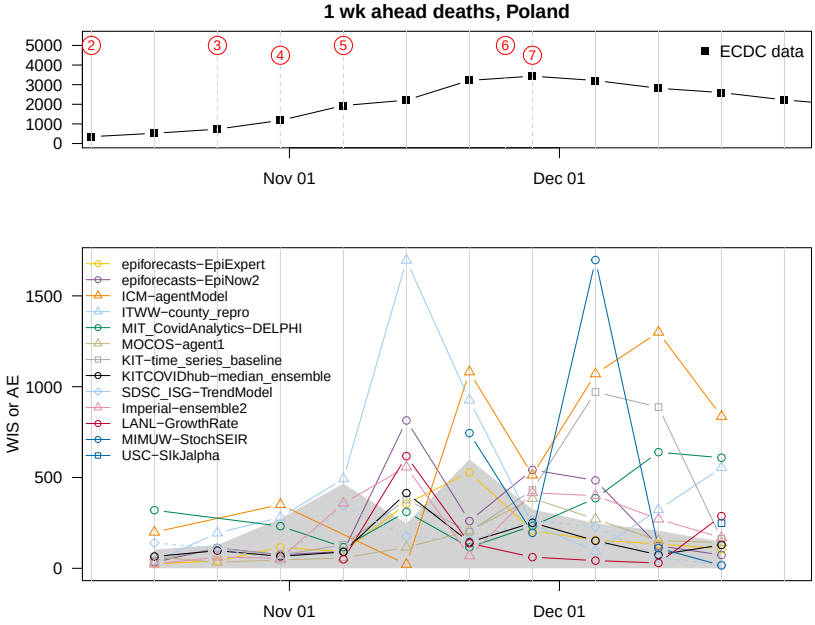


Figure B.8.: WIS value or absolute error (of deterministic models) over time, for one-week-ahead death forecasts for Poland. Numbers in circles represent events explained in Figure 3.1. As in Figure 3.7, the upper border of the light grey area represents the performance of the naïve baseline model `KIT-baseline`.

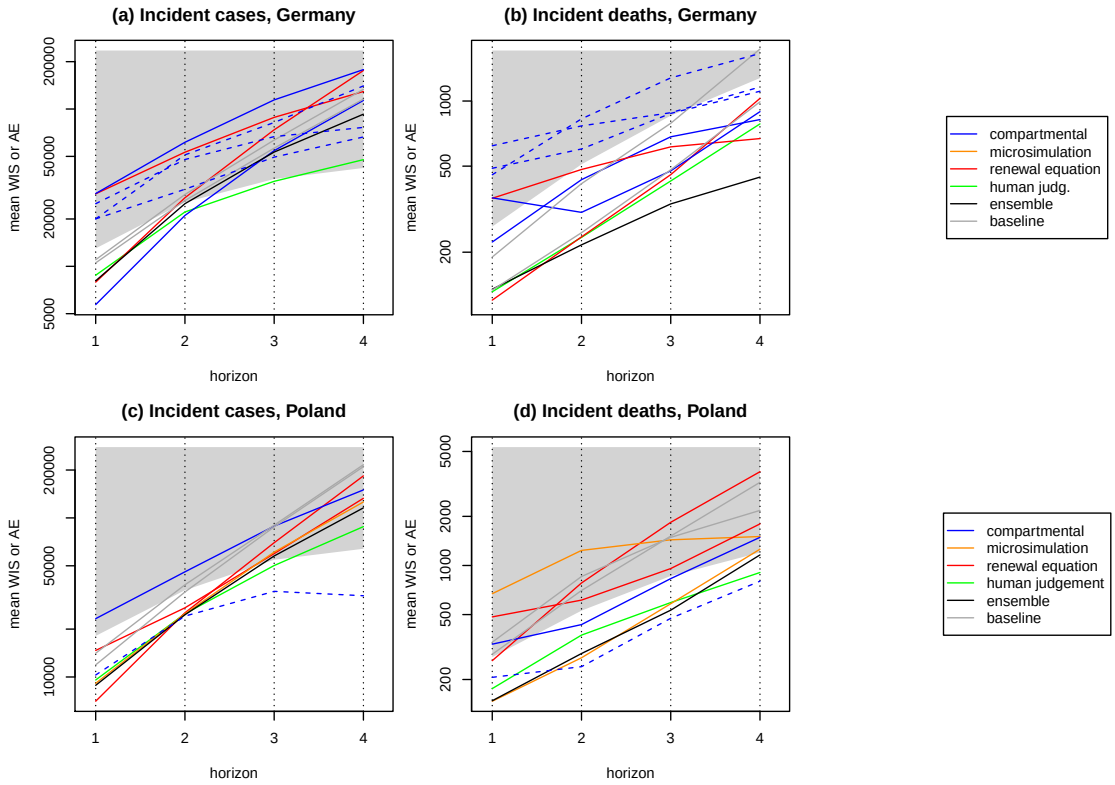


Figure B.9.: Mean WIS or AE values by target and forecast horizon as in Figure 3.7. Lines are coloured according to the model categories introduced in Table 3.3. No clear patterns emerge apart from the fact that the ensemble model shows good relative performance for death forecasts.

B.8. Results for three- and four-week-ahead forecasts

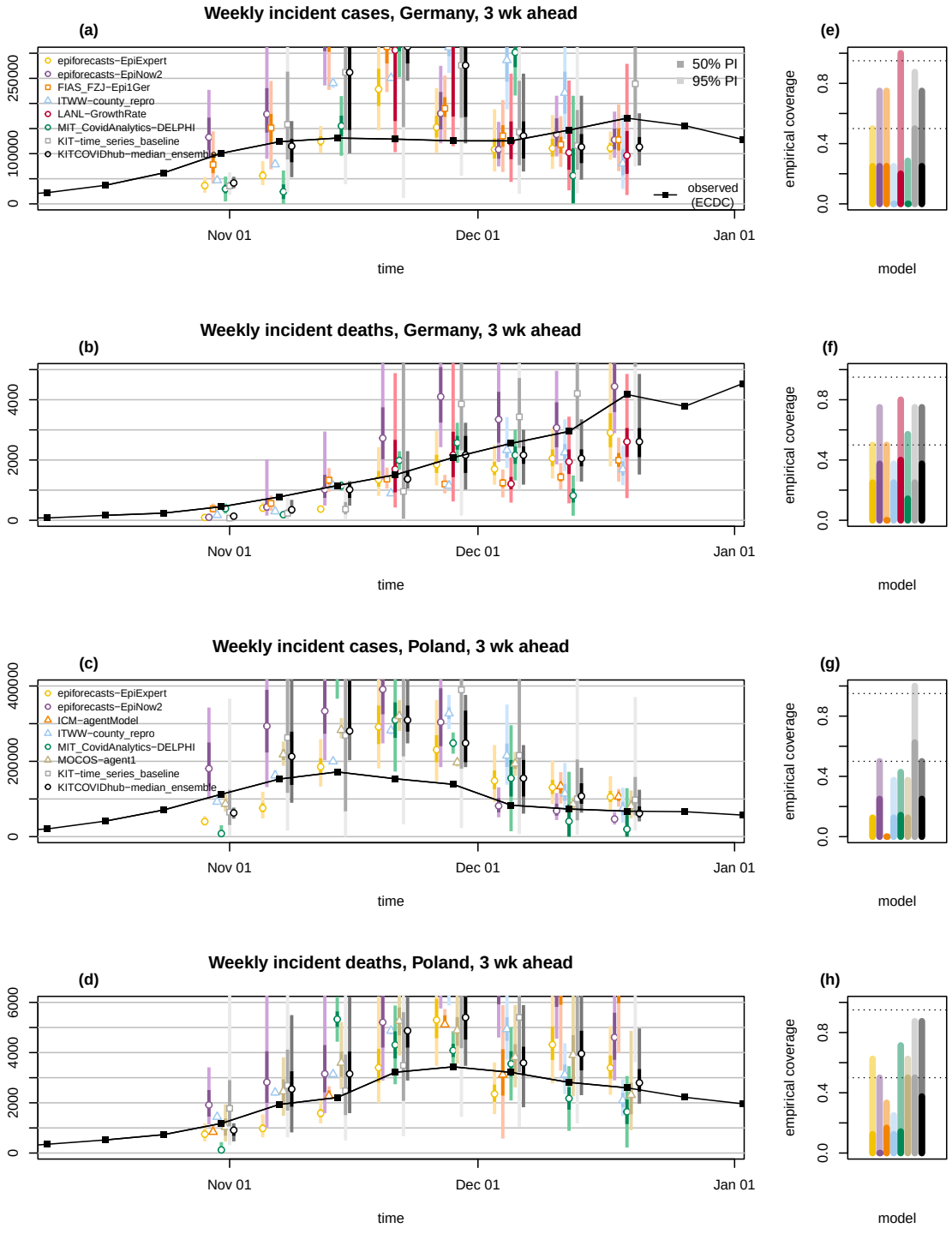


Figure B.10.: Three-week-ahead forecasts. Three-week-ahead forecasts of incident cases and deaths in Germany (**a, b**) and Poland (**c, d**). Displayed are predictive medians, 50% and 95% prediction intervals (PIs). Coverage plots (**e**–**h**) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

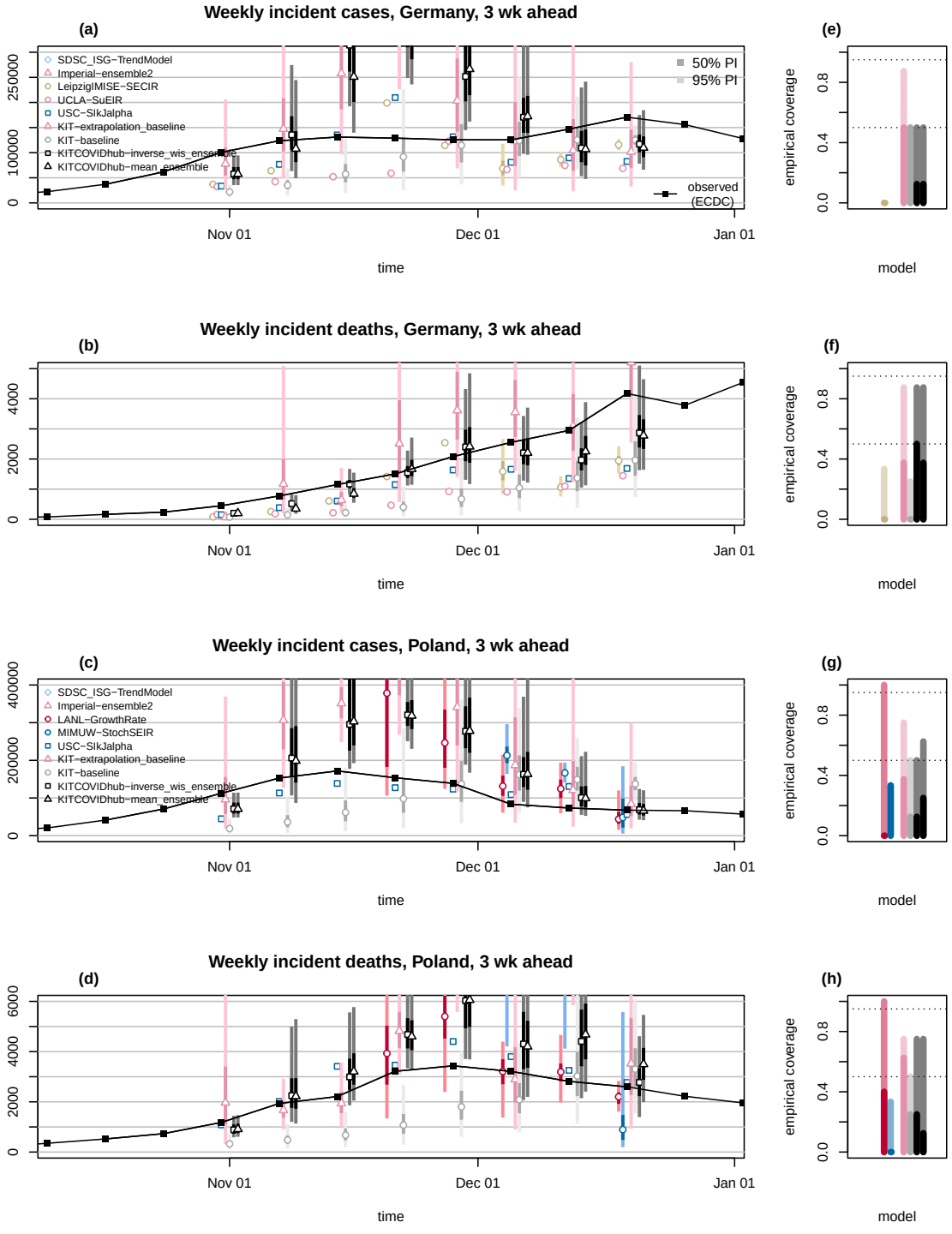


Figure B.11.: Additional three-week-ahead forecasts. Three-week-ahead forecasts of incident cases and deaths in Germany (**a, b**) and Poland (**c, d**). Displayed are predictive medians, 50% and 95% prediction intervals (PIs) for models not shown in Figure B.10. Coverage plots (**e**–**h**) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

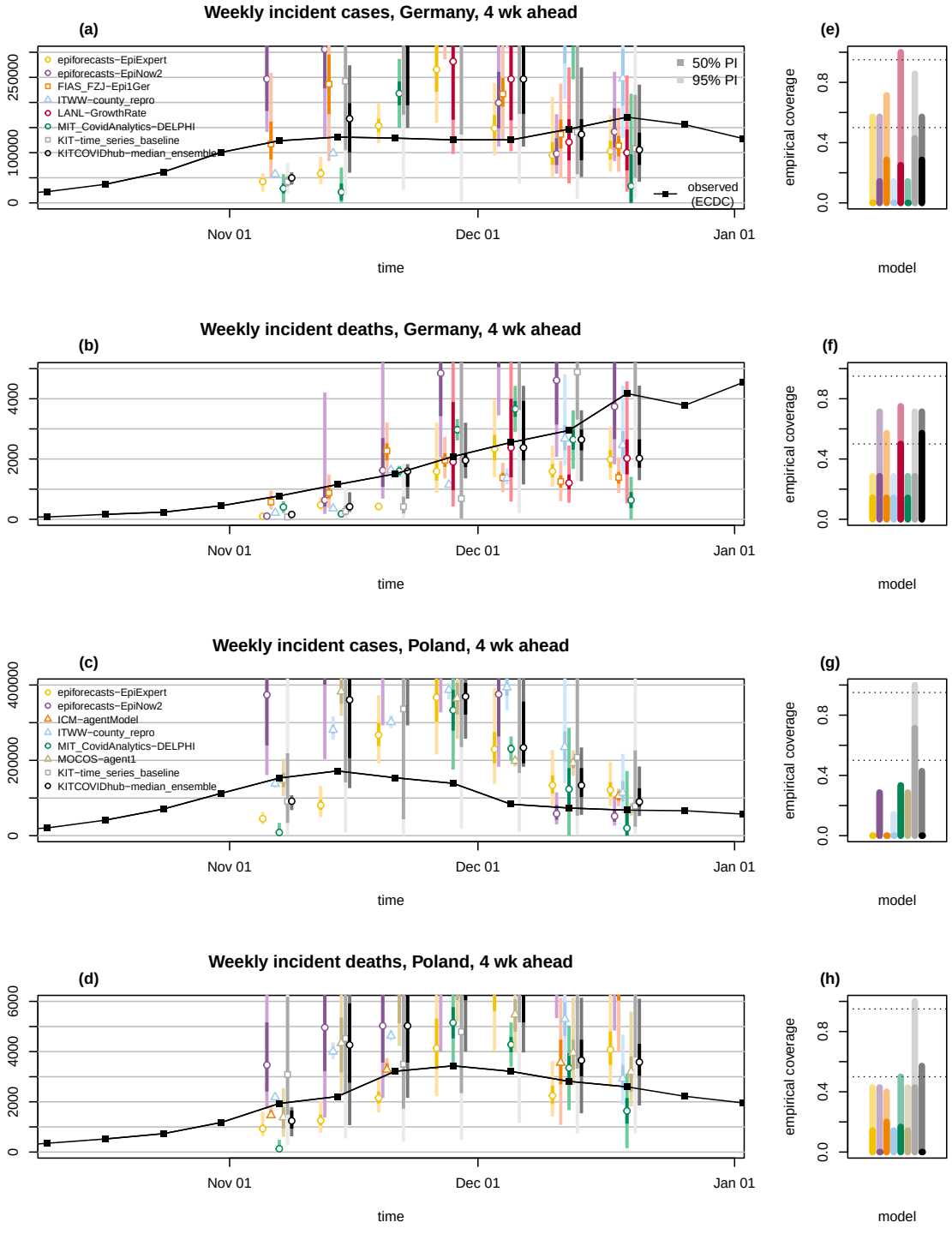


Figure B.12.: Four-week-ahead forecasts. Four-week-ahead forecasts of incident cases and deaths in Germany (**a, b**) and Poland (**c, d**). Displayed are predictive medians, 50% and 95% prediction intervals (PIs). Coverage plots (**e**–**h**) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

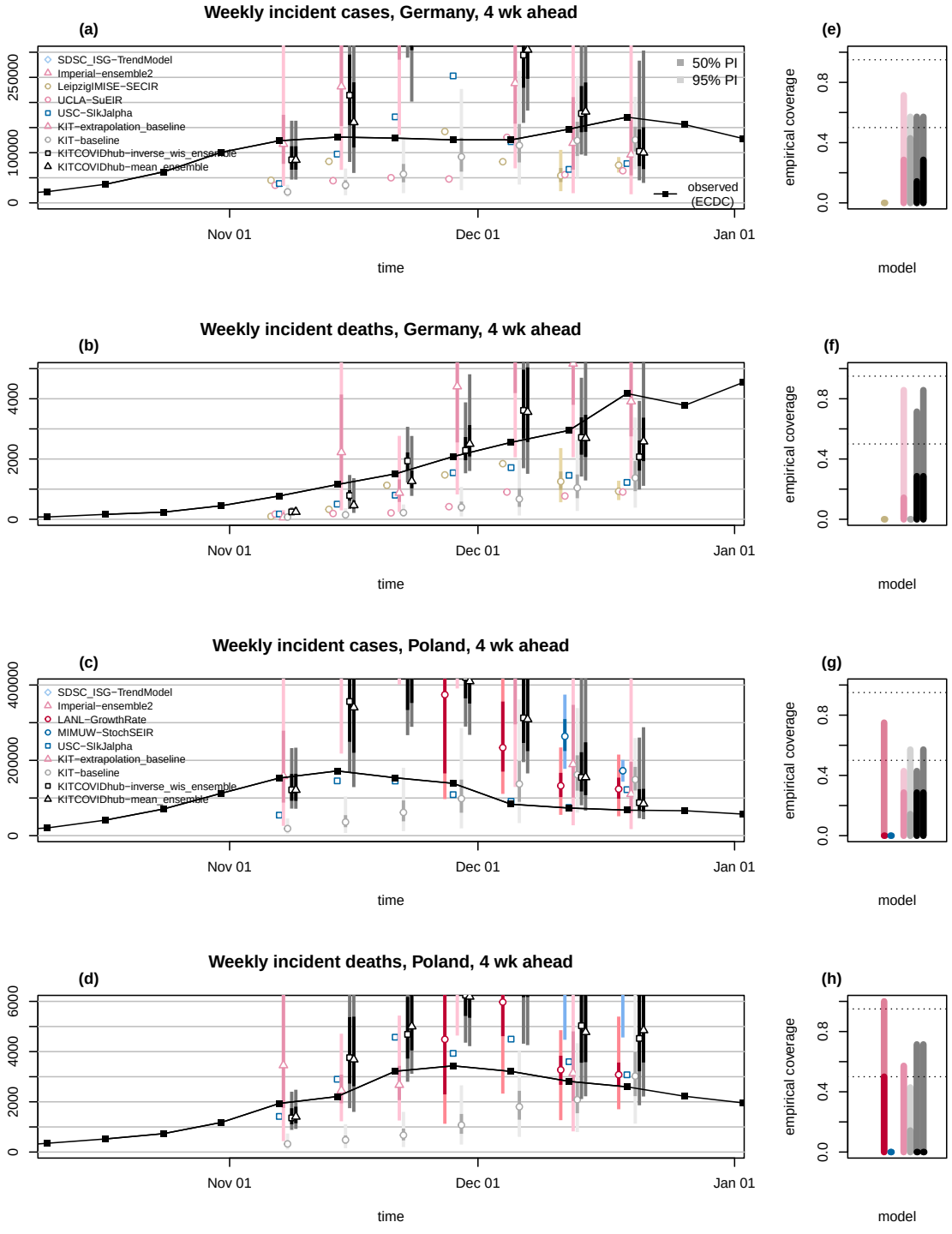


Figure B.13.: Additional four-week-ahead forecasts. Four-week-ahead forecasts of incident cases and deaths in Germany (**a, b**) and Poland (**c, d**). Displayed are predictive medians, 50% and 95% prediction intervals (PIs) for models not shown in Figure B.12. Coverage plots (**e**–**h**) show the empirical coverage of 95% (light) and 50% (dark) prediction intervals.

Table B.2.: Detailed summary of forecast evaluation for Germany (based on JHU data)

Germany, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	13,351	9,470	3/10	7/10	33,588	24,996	2/9	3/9	13,351	9,470	3/10	7/10	46,506	34,234	2/9	4/9
epiforecasts-EpiNow2	9,879	6,413	6/10	8/10	35,103	26,003	4/9	6/9	9,879	6,413	6/10	8/10	43,106	31,285	5/9	7/9
FIAS_FZJ-Epi1Ger	7,377	4,720	5/10	10/10	29,886	20,438	4/9	7/9	14,775	9,326	3/10	7/10	38,614	27,427	4/9	6/9
ITWW-county_repro	34,825	29,077	0/10	2/10	63,812	52,478	0/9	2/9	34,476	28,729	0/10	2/10	100,873	83,619	0/9	1/9
LANL-GrowthRate	38,679*	22,973*	4/7	7/7	79,788*	43,165*	2/6	6/6	39,734	27,405	4/10	7/10	120,334	81,370	2/9	5/9
LeipzigIMISE-SECIR	20,064		1/5	3/5	51,576		0/4	1/4	36,089	31,838	0/10	1/10	93,828	84,209	1/9	2/9
MIT_CovidAnalytics-DELPHI	40,959*	29,499*	2/8	4/8	82,336*	63,549*	0/7	2/7								
SDSC_ISG-TrendModel	14,173								14,173							
UCLA-SuEIR	28,331				50,970				28,331				76,523			
USC-SIkJalpha	21,935		1/1	1/1	34,150				23,474		1/1	1/1	53,480			
KIT-baseline	21,980	15,006	4/10	8/10	35,913	28,029	3/9	5/9	21,980	15,006	4/10	8/10	53,840	42,428	3/9	5/9
KIT-extrapolation_baseline	12,851	10,487	7/10	10/10	37,194	26,323	5/9	7/9	12,851	10,487	7/10	10/10	47,861	34,846	7/9	7/9
KIT-time_series_baseline	14,910	10,950	6/10	9/10	43,955	28,311	5/9	8/9	14,910	10,950	6/10	9/10	59,721	38,684	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	14,205	8,992	3/10	9/10	42,759	27,897	1/9	6/9	13,525	8,813	4/10	8/10	53,689	34,975	2/9	7/9
KITCOVIDhub-mean_ensemble	17,484	10,712	3/10	8/10	42,910	27,776	2/9	6/9	16,171	10,397	3/10	7/10	57,842	37,415	1/9	6/9
KITCOVIDhub-median_ensemble	12,271	8,001	4/10	9/10	38,316	25,784	3/9	7/9	13,277	8,962	4/10	7/10	50,155	34,504	2/9	6/9

Germany, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	236	156	4/10	7/10	390	268	3/9	6/9	236	156	4/10	7/10	545	350	2/9	7/9
epiforecasts-EpiNow2	159	108	5/10	8/10	357	218	4/9	7/9	159	108	5/10	8/10	485	296	5/9	7/9
FIAS_FZJ-Epi1Ger	307	268	2/10	4/10	579	484	1/9	3/9	244	206	0/10	3/10	779	655	2/9	4/9
Imperial-ensemble2	306	225	2/10	5/10					303	222	2/10	5/10				
ITWW-county_repro	420	403	0/10	2/10	594	536	0/9	1/9	419	402	0/10	2/10	924	852	0/9	2/9
LANL-GrowthRate	234*	148*	4/7	7/7	500*	353*	2/6	5/6	218	149	4/10	7/10	658	493	3/9	5/9
LeipzigIMISE-SECIR	653		0/5	1/5	815		1/4	1/4	1,199	1,024	0/10	1/10	2,247	1,867	0/9	1/9
MIT_CovidAnalytics-DELPHI	499*	378*	2/8	3/8	401*	313*	1/7	5/7	473*	353*	2/8	3/8	635*	533*	1/7	3/7
SDSC_ISG-TrendModel	409								409							
UCLA-SuEIR	508				884				508				1,281			
USC-SIkJalpha	540		0/1	0/1	657				552		0/1	0/1	1,066			
KIT-baseline	531	291	0/10	9/10	892	554	0/9	5/9	531	291	0/10	9/10	1,258	789	0/9	5/9
KIT-extrapolation_baseline	181	135	7/10	9/10	385	244	5/9	8/9	181	135	7/10	9/10	483	319	6/9	8/9
KIT-time_series_baseline	213	179	6/10	9/10	592	402	4/9	8/9	213	179	6/10	9/10	803	547	4/9	8/9
KITCOVIDhub-inverse_wis_ensemble	218	142	2/10	8/10	302	178	1/9	8/9	214	136	2/10	8/10	474	279	1/9	6/9
KITCOVIDhub-mean_ensemble	256	168	1/10	9/10	345	206	2/9	8/9	267	177	2/10	9/10	544	324	1/9	8/9
KITCOVIDhub-median_ensemble	251	165	2/10	8/10	381	251	2/9	7/9	242	159	3/10	8/10	543	330	2/9	8/9

Table B.3.: Detailed summary of forecast evaluation for Poland (based on JHU data)

Poland, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	16,105	11,390	4/10	8/10	41,186	27,900	1/9	5/9	16,128	11,414	4/10	8/10	59,079	38,550	1/9	6/9
epiforecasts-EpiNow2	9,147	6,587	7/10	7/10	36,477	24,085	2/9	6/9	9,170	6,594	7/10	7/10	42,849	28,105	6/9	7/9
ICM-agentModel			2/4	4/4			0/3	1/3			1/4	3/4			0/3	2/3
ITWW-county_repro	19,441	16,205	1/10	4/10	37,152	30,879	2/9	2/9	18,519	15,278	1/10	3/10	57,467	46,975	1/9	3/9
LANL-GrowthRate	13,494*	8,719*	5/7	7/7	48,693*	27,686*	1/6	6/6	15,205	10,151	5/10	8/10	66,280	39,924	2/9	8/9
MIMUW-StochSEIR			3/5	4/5			2/4	2/4			2/5	4/5			1/4	2/4
MIT_CovidAnalytics-DELPHI	30,505*	22,627*	3/9	5/9	61,303*	46,360*	1/8	4/8								
MOCOS-agent1	15,732	11,468	1/10	4/10	32,235	26,111	1/9	3/9	15,732	11,468	1/10	4/10	46,214	35,642	1/9	3/9
SDSC_ISG-TrendModel	10,817								10,888							
USC-SIkJalpha	13,544		0/1	1/1	27,115				16,754		0/1	1/1	41,949			
KIT-baseline	31,605	20,001	5/10	9/10	55,931	37,396	2/9	6/9	31,676	20,036	5/10	9/10	87,597	59,162	2/9	6/9
KIT-extrapolation_baseline	18,333	11,754	7/10	10/10	55,685	34,091	3/9	8/9	18,311	11,752	7/10	10/10	77,269	45,119	3/9	9/9
KIT-time_series_baseline	22,502	14,455	5/10	10/10	60,704	38,643	4/9	9/9	22,480	14,452	5/10	10/10	85,192	53,310	4/9	9/9
KITCOVIDhub-inverse_wis_ensemble	14,191	9,103	5/10	8/10	37,174	25,943	3/9	6/9	14,325	8,838	5/10	8/10	50,096	33,472	2/9	6/9
KITCOVIDhub-mean_ensemble	13,849	8,879	4/10	9/10	37,831	24,819	2/9	6/9	14,511	8,727	5/10	8/10	50,731	32,216	2/9	7/9
KITCOVIDhub-median_ensemble	15,236	9,529	6/10	9/10	40,453	25,715	2/9	6/9	16,541	10,105	4/10	8/10	55,827	34,268	1/9	5/9

Poland, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	303	186	3/10	9/10	625	399	2/9	7/9	303	186	3/10	9/10	911	571	2/9	8/9
epiforecasts-EpiNow2	343	240	5/10	7/10	1,066	758	3/9	5/9	344	240	5/10	7/10	1,439	1,003	3/9	5/9
ICM-agentModel	808*	715*	1/8	2/8	1,921*	1,272*	0/7	2/7	1,234*	770*	0/8	3/8	3,000*	1,904*	0/7	4/7
Imperial-ensemble2	379	229	6/10	7/10					351	200	6/10	7/10				
ITWW-county_repro	465	428	1/10	1/10	652	560	0/9	2/9	458	422	2/10	2/10	1,099	968	0/9	3/9
LANL-GrowthRate	236*	158*	4/7	7/7	383*	235*	4/6	6/6	236	154	4/10	9/10	655	420	3/9	8/9
MIMUW-StochSEIR			2/5	4/5			1/4	2/4			2/5	4/5			0/4	4/4
MIT_CovidAnalytics-DELPHI	467*	299*	2/9	7/9	621*	409*	1/8	6/8	552*	386*	2/9	7/9	990*	705*	1/8	4/8
MOCOS-agent1	205	153	8/10	10/10	393	253	5/9	8/9	205	153	8/10	10/10	533	359	7/9	9/9
SDSC_ISG-TrendModel	180								179							
USC-SIkJalpha	202		0/1	1/1	212				252		0/1	0/1	283			
KIT-baseline	504	305	5/10	10/10	882	578	2/9	6/9	503	304	5/10	10/10	1,365	896	2/9	5/9
KIT-extrapolation_baseline	412	274	6/10	8/10	995	700	5/9	7/9	411	274	6/10	8/10	1,422	974	4/9	7/9
KIT-time_series_baseline	528	333	8/10	10/10	1,343	853	5/9	8/9	529	333	8/10	10/10	1,909	1,206	5/9	8/9
KITCOVIDhub-inverse_wis_ensemble	207	138	7/10	10/10	476	300	4/9	9/9	231	151	7/10	10/10	683	433	4/9	8/9
KITCOVIDhub-mean_ensemble	227	147	7/10	10/10	558	349	4/9	9/9	250	158	6/10	10/10	793	500	4/9	9/9
KITCOVIDhub-median_ensemble	195	134	6/10	10/10	460	278	4/9	9/9	215	147	6/10	10/10	686	433	4/9	8/9

Table B.4.: Summary of forecast evaluation for ensembles without plausibility checks of members (based on ECDC data)

Germany, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
KITCOVIDhub-inverse_wis_ensemble_all	13,431	8,835	4/10	9/10	39,275	24,810	1/9	6/9	28,345	18,730	1/10	7/10	55,290	35,321	2/9	7/9
KITCOVIDhub-mean_ensemble_all	15,554	9,848	4/10	9/10	40,120	24,956	1/9	6/9	16,068	10,397	4/10	7/10	54,550	34,516	2/9	6/9
KITCOVIDhub-median_ensemble_all	11,240	7,959	6/10	9/10	36,823	23,838	3/9	7/9	13,511	9,593	6/10	7/10	51,242	35,038	2/9	6/9

Germany, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
KITCOVIDhub-inverse_wis_ensemble_all	177	109	6/10	9/10	234	144	4/9	8/9	845	665	0/10	8/10	1,065	742	1/9	8/9
KITCOVIDhub-mean_ensemble_all	183	124	6/10	8/10	263	162	4/9	8/9	236	157	4/10	6/10	472	291	3/9	6/9
KITCOVIDhub-median_ensemble_all	185	129	6/10	8/10	332	217	3/9	7/9	196	132	4/10	7/10	434	270	3/9	7/9

Poland, cases

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
KITCOVIDhub-inverse_wis_ensemble_all	12,100	8,065	4/10	9/10	36,692	23,049	3/9	7/9	11,951	7,300	5/10	10/10	44,256	28,051	2/9	8/9
KITCOVIDhub-mean_ensemble_all	11,788	7,847	4/10	9/10	37,031	22,548	2/9	8/9	11,649	7,076	5/10	10/10	43,910	27,625	3/9	9/9
KITCOVIDhub-median_ensemble_all	13,597	8,632	5/10	9/10	39,156	23,726	2/9	7/9	13,365	8,415	5/10	9/10	48,278	28,472	3/9	7/9

Poland, deaths

Model	1 wk ahead inc				2 wk ahead inc				1 wk ahead cum				2 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
KITCOVIDhub-inverse_wis_ensemble_all	188	141	6/10	10/10	467	295	5/9	9/9	384	241	2/10	9/10	656	406	5/9	9/9
KITCOVIDhub-mean_ensemble_all	204	147	7/10	9/10	593	353	4/9	9/9	220	154	7/10	9/10	802	485	3/9	9/9
KITCOVIDhub-median_ensemble_all	202	138	6/10	10/10	428	272	6/9	9/9	194	135	8/10	10/10	620	404	6/9	9/9

Table B.5.: Detailed summary of forecast evaluation for Germany, 3 and 4 weeks ahead (based on ECDC data)

Germany, cases

Model	3 wk ahead inc				4 wk ahead inc				3 wk ahead cum				4 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	47,162	34,704	2/8	4/8	65,906	47,640	0/7	4/7	85,859	60,488	2/8	4/8	143,536	101,205	2/7	4/7
epiforecasts-EpiNow2	98,948	73,773	2/8	6/8	235,461	175,688	1/7	4/7	148,866	109,915	3/8	6/8	394,730	293,483	1/7	5/7
FIAS_FZJ-Epi1Ger	74,550	54,378	2/8	6/8	149,451	113,410	2/7	5/7	116,086	84,710	4/8	6/8	271,463	203,058	2/7	4/7
ITWW-county_repro	105,132	88,509	0/8	2/8	147,570	129,219	0/7	1/7	214,867	180,139	0/8	1/8	369,098	319,826	0/7	1/7
LANL-GrowthRate			1/5	5/5			1/4	4/4	218,206	152,213	2/8	5/8	328,115	246,869	2/7	4/7
LeipzigIMISE-SECIR	82,104		0/3	0/3	140,275		0/2	0/2	171,325	153,731	1/8	2/8	296,564	267,098	2/7	2/7
MIT_CovidAnalytics-DELPHI	139,142*	114,298*	0/7	2/7	210,370	178,016	0/7	1/7								
SDSC_ISG-TrendModel																
UCLA-SuEIR	66,768				76,415				132,374				198,546			
USC-SIkJalpha	49,446				66,436				96,902				140,716			
KIT-baseline	44,706	35,891	4/8	4/8	54,563	42,192	3/7	4/7	85,838	70,756	3/8	4/8	136,280	115,148	2/7	4/7
KIT-extrapolation_baseline	82,243	56,387	4/8	7/8	165,710	115,568	2/7	5/7	125,152	88,749	4/8	7/8	291,137	207,420	4/7	6/7
KIT-time_series_baseline	91,848	63,486	4/8	7/8	162,293	133,341	3/7	6/7	154,663	103,058	3/8	7/8	320,671	239,590	3/7	6/7
KITCOVIDhub-inverse_wis_ensemble	90,487	65,207	1/8	4/8	171,254	126,851	1/7	4/7	136,583	93,984	3/8	5/8	278,362	206,835	3/7	5/7
KITCOVIDhub-mean_ensemble	82,294	57,308	1/8	4/8	138,225	104,078	2/7	4/7	137,562	92,565	3/8	5/8	262,811	190,439	3/7	4/7
KITCOVIDhub-median_ensemble	79,238	53,265	2/8	6/8	129,400	92,703	2/7	4/7	125,538	89,344	2/8	5/8	252,946	191,986	3/7	4/7

Germany, deaths

Model	3 wk ahead inc				4 wk ahead inc				3 wk ahead cum				4 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	615	427	2/8	4/8	955	784	1/7	2/7	981	655	2/8	5/8	1,850	1,379	2/7	3/7
epiforecasts-EpiNow2	655	457	3/8	6/8	1,511	1,029	2/7	5/7	1,044	709	3/8	6/8	2,532	1,735	2/7	5/7
FIAS_FZJ-Epi1Ger	811	683	0/8	4/8	1,003	821	1/7	4/7	1,467	1,249	1/8	4/8	2,153	1,785	0/7	4/7
Imperial-ensemble2																
ITWW-county_repro	713	614	2/8	3/8	798	671	1/7	2/7	1,345	1,174	2/8	2/8	1,740	1,487	1/7	3/7
LANL-GrowthRate			2/5	4/5			2/4	3/4	1,364	1,098	2/8	4/8	2,238	2,006	2/7	3/7
LeipzigIMISE-SECIR	883		0/3	1/3	1,162		0/2	0/2	2,722	2,283	1/8	2/8	2,780	2,388	1/7	2/7
MIT_CovidAnalytics-DELPHI	593*	475*	1/7	4/7	1,039	893	1/7	2/7	1,211*	1,054*	0/7	1/7	2,201	2,018	0/7	0/7
SDSC_ISG-TrendModel																
UCLA-SuEIR	1,279				1,659				2,273				3,615			
USC-SIkJalpha	877				1,110				1,635				2,344			
KIT-baseline	1,218	855	0/8	2/8	1,608	1,277	0/7	0/7	2,186	1,584	0/8	0/8	3,538	2,947	0/7	0/7
KIT-extrapolation_baseline	752	481	3/8	7/8	1,526	992	1/7	6/7	1,252	799	3/8	7/8	2,666	1,731	1/7	6/7
KIT-time_series_baseline	1,092	789	2/8	6/8	1,712	1,739	2/7	5/7	1,727	1,259	3/8	7/8	2,934	2,802	2/7	5/7
KITCOVIDhub-inverse_wis_ensemble	440	274	4/8	7/8	704	471	2/7	5/7	810	499	2/8	6/8	1,259	881	4/7	4/7
KITCOVIDhub-mean_ensemble	488	297	3/8	7/8	676	433	2/7	6/7	880	557	3/8	6/8	1,301	837	3/7	5/7
KITCOVIDhub-median_ensemble	493	334	3/8	6/8	599	445	4/7	5/7	897	590	3/8	6/8	1,310	978	3/7	5/7

Table B.6.: Detailed summary of forecast evaluation for Poland, 3 and 4 weeks ahead (based on ECDC data)

Poland, cases

Model	3 wk ahead inc				4 wk ahead inc				3 wk ahead cum				4 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	69,036	50,186	1/8	1/8	114,361	88,135	0/7	0/7	123,420	83,370	1/8	5/8	236,085	163,096	0/7	2/7
epiforecasts-EpiNow2	100,091	69,829	2/8	4/8	257,145	183,720	2/7	2/7	146,121	96,446	3/8	6/8	422,386	291,126	2/7	4/7
ICM-agentModel			0/2	0/2			0/1	0/1			0/2	1/2			0/1	0/1
ITWW-county_repro	69,096	59,629	1/8	3/8	148,272	132,809	0/7	1/7	115,872	95,988	1/8	4/8	271,133	233,566	0/7	1/7
LANL-GrowthRate			0/5	5/5			0/4	3/4	147,981	85,560	1/8	7/8	278,420	163,238	1/7	5/7
MIMUW-StochSEIR			1/3	1/3			0/2	0/2			1/3	1/3			0/2	0/2
MIT_CovidAnalytics-DELPHI	113,802*	88,815*	1/7	3/7	177,706*	149,967*	2/6	2/6								
MOCOS-agent1	70,362	60,944	1/8	3/8	140,691	125,932	2/7	2/7	116,589	94,704	0/8	2/8	263,653	226,693	0/7	2/7
SDSC_ISG-TrendModel																
USC-SIkJalpha	34,543				32,410				71,142				110,078			
KIT-baseline	75,183	54,568	1/8	4/8	89,395	63,975	1/7	4/7	159,350	118,193	1/8	3/8	238,036	183,364	1/7	3/7
KIT-extrapolation_baseline	125,846	87,821	3/8	6/8	278,666	210,038	2/7	3/7	206,056	135,329	3/8	6/8	505,952	360,439	2/7	4/7
KIT-time_series_baseline	123,714	90,048	5/8	8/8	228,443	215,974	5/7	7/7	217,544	145,712	5/8	8/8	476,907	376,333	5/7	7/7
KITCOVIDhub-inverse_wis_ensemble	78,949	57,514	1/8	4/8	160,008	129,539	2/7	3/7	115,211	82,871	2/8	5/8	266,961	197,054	1/7	5/7
KITCOVIDhub-mean_ensemble	78,772	56,181	2/8	5/8	156,341	125,110	2/7	4/7	115,204	80,176	2/8	6/8	265,775	191,339	1/7	5/7
KITCOVIDhub-median_ensemble	74,351	57,640	2/8	4/8	144,957	115,659	0/7	3/7	126,339	77,669	1/8	5/8	275,236	171,357	1/7	4/7

Poland, deaths

Model	3 wk ahead inc				4 wk ahead inc				3 wk ahead cum				4 wk ahead cum			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	901	592	1/8	5/8	1,329	906	1/7	3/7	1,828	1,135	1/8	6/8	3,171	2,035	1/7	4/7
epiforecasts-EpiNow2	2,642	1,840	0/8	4/8	5,317	3,757	0/7	3/7	4,316	2,977	1/8	4/8	9,641	6,722	0/7	4/7
ICM-agentModel	1,854*	1,437*	1/6	2/6	1,850*	1,505*	1/5	2/5	4,376*	2,847*	1/6	2/6	4,584*	3,275*	1/5	3/5
Imperial-ensemble2																
ITWW-county_repro	1,113	958	1/8	2/8	2,016	1,799	1/7	1/7	2,375	2,080	0/8	2/8	4,592	4,103	0/7	1/7
LANL-GrowthRate			2/5	5/5			2/4	4/4	1,517	938	4/8	7/8	2,984	1,956	3/7	6/7
MIMUW-StochSEIR			0/3	1/3			0/2	0/2			0/3	1/3			0/2	0/2
MIT_CovidAnalytics-DELPHI	1,122*	831*	1/7	5/7	1,851*	1,488*	1/6	3/6	2,076*	1,728*	1/7	1/7	3,899*	3,484*	0/6	1/6
MOCOS-agent1	940	583	4/8	5/8	1,883	1,259	1/7	3/7	1,450	890	5/8	7/8	3,497	2,072	3/7	4/7
SDSC_ISG-TrendModel																
USC-SIkJalpha	474				801				597				1,445			
KIT-baseline	1,208	862	2/8	4/8	1,544	1,201	1/7	3/7	2,368	1,768	1/8	4/8	3,986	3,226	1/7	2/7
KIT-extrapolation_baseline	1,995	1,514	5/8	6/8	4,093	3,234	4/7	4/7	3,537	2,561	4/8	6/8	7,690	5,969	3/7	4/7
KIT-time_series_baseline	2,235	1,487	4/8	7/8	3,084	2,176	3/7	7/7	3,799	2,505	5/8	7/8	5,787	3,872	4/7	7/7
KITCOVIDhub-inverse_wis_ensemble	1,039	649	2/8	6/8	2,205	1,447	0/7	5/7	1,668	1,069	4/8	8/8	3,850	2,467	2/7	5/7
KITCOVIDhub-mean_ensemble	1,165	710	1/8	6/8	2,232	1,461	0/7	5/7	1,978	1,242	3/8	6/8	4,109	2,653	1/7	5/7
KITCOVIDhub-median_ensemble	895	532	3/8	7/8	1,871	1,161	0/7	4/7	1,645	1,030	4/8	7/8	3,656	2,199	2/7	4/7

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

4.1. 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 (Ray et al., 2020; Cramer et al., 2022b) and Scenario Modelling Hubs (Borchering et al., 2021), supported by the US Centers for Disease Control and Prevention, as well as the more recent European COVID-19 Forecast Hub (Sherratt et al., 2023), 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 (McGowan et al., 2019), the DARPA Chikungunya Challenge (Del Valle et al., 2018) or the Dengue Forecasting Project (Johansson et al., 2019) 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 (Nature Publishing Group, 2020). It has been highlighted that such modelling studies should be prospective (Arık et al., 2021) and ideally follow pre-registered protocols (Dirnagl, 2021) 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 (Bracher et al., 2020) 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 the national health authorities (RKI, Robert Koch Institute (2023) and MZ, Polish Ministry of Health (2022); the primary data source) or the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE (2022) and Dong et al. (2020)). 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 4.1. 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 (Bracher et al., 2021b). 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 4.1 shows the estimated weekly percentages of all cases which were due to the B.1.1.7 variant in Germany (Robert Koch Institute, 2021a) and Poland (MI2 Data Lab, Warsaw University of Technology, 2021; GISAID Initiative, 2021) 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 (Hale et al., 2021). 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 the 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). Note that all these data are publicly available via the respective public health agencies and their use does not require ethical approval.

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.

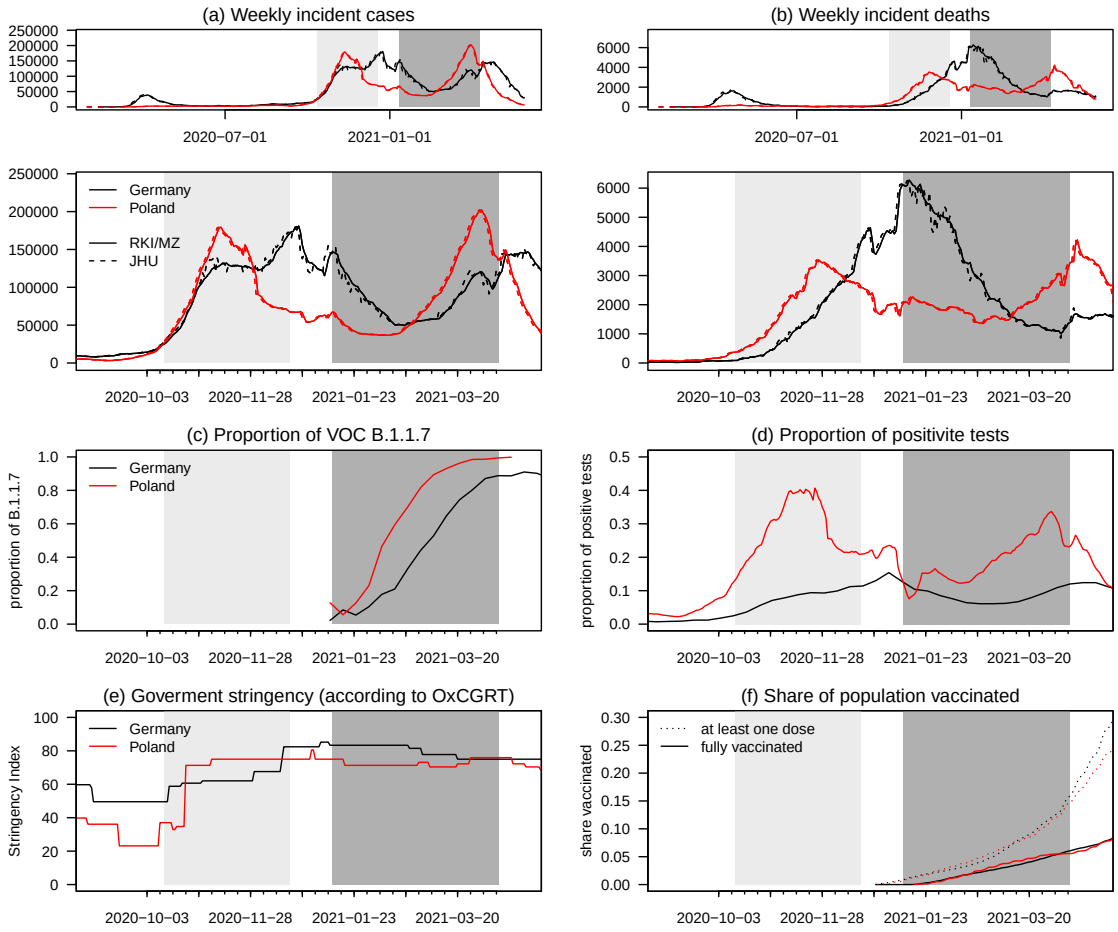


Figure 4.1.: Overview of relevant epidemiological time series. Reported cases (a) and deaths (b) in Germany (black) and Poland (red) according to Robert Koch Institute, the Polish Ministry of Health (MZ; solid lines) and Johns Hopkins CSSE (dashed). Additional panels show (c) the share of cases due to the B.1.1.7 (Alpha) variant, (d) the proportion of all performed PCR tests that 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. (2021b).

4.2. Methods

The methods described in the following are largely identical to those in the first part of our study (Bracher et al., 2021b), but are presented to ensure self-containedness of the present work.

4.2.1. 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 the Robert Koch Institute and the Polish Ministry of Health. All data streams were aggregated by the time of appearance in national data, see also Supplementary Note 4 of Bracher et al. (2021b). 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 (Bracher et al., 2021b) 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.

4.2.2. 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 (Gneiting and Raftery, 2007). 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

$$\mathrm{IS}_\alpha(F, y) = (u - l) \; + \; \frac{2}{\alpha} \times (l - y) \times \chi(y < l) \; + \; \frac{2}{\alpha} \times (y - u) \times \chi(y > u), \qquad (4.1)$$

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; Bracher et al. (2021a)). 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

$$\text{WIS}(F, y) = \frac{1}{2N+1} \times \left(|y - m| \; + \; \sum_{k=1}^{N} (\alpha_k \times \text{IS}_{\alpha_k}(F, y)) \right), \tag{4.2}$$

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; Gneiting and Raftery (2007)) and generalizes the absolute error to probabilistic forecasts. 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.

4.2.3. 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. (2021b) for detailed descriptions. Table 4.1 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 Appendix C.1. All forecast data

Table 4.1.: Forecast models contributed by independent external research teams.

Category	Model	NPI	Test	Variants	Age	DE	PL	Regional	Truth	Pr
Agent-based	ICM-agentModel (Rakowski et al., 2010)	✓	✓	✓	✓		✓		MZ	✓
	MOCOS-agent1 (Adamik et al., 2020)	✓	✓	✓	✓		✓		JHU	✓
Compartment	CovidAnalytics-DELPHI (Li et al., 2020)	✓				✓	✓		JHU	✓
	FIAS_FZJ-Epi1Ger (Barbarossa et al., 2020)					✓		✓	RKI	✓
	itwm-dSEIR				✓	✓			RKI	✓
	Karlen-pypm (Karlen, 2020)			✓		✓		✓	RKI	✓
	LeipzigIMISE-SECIR (Kheifetz et al., 2021)	✓	✓	✓		✓		(✓)	RKI	✓
	MIMUW-StochSEIR						✓		JHU	✓
	USC-SIkJalpha (Srivastava et al., 2020)					✓	✓	✓	RKI/MZ	✓
Growth rate/ renewal eq.	epiforecasts-EpiNow2 (Abbott et al., 2020b)					✓	✓	✓	RKI/MZ	✓
	SDSC_ISG-TrendModel (Krymova et al., 2022)					✓	✓		JHU	
	ITWW-county_repro (Burgard et al., 2021)				✓	✓	✓	✓	RKI/MZ	✓
	LANL-GrowthRate (Castro et al., 2021)					✓	✓		JHU	✓
Human judgement	epiforecasts-EpiExpert (Bosse et al., 2021)	(✓)	(✓)	(✓)	(✓)	✓	✓		RKI/MZ	✓
Forecast ensemble	Imperial-ensemble2 (Bhatia et al., 2021)					✓	✓		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. (2021b), Supplementary Note 3 and in the Supplementary Methods (Section 1) of this article.

produced by teams was made available under open licences. They do not contain any personal data so that no ethics approval was required for their use.

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 (Cramer et al., 2022b), 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. (2021b), Supplementary Note 2). KIT-baseline 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 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 is an exponential smoothing time series model with multiplicative errors as used by Petropoulos and Makridakis (2020) to predict COVID-19 cases and deaths. It is implemented using the R package forecast, version 8.12 (Hyndman, 2021).

As a further external comparison we added publicly available death forecasts by the Institute for Health Metrics and Evaluation (IHME (2021), University of Washington; available under the CC BY-NC 4.0 license). Here, we always used the most recent prediction available on a given forecast date.

4.2.4. Forecast ensembles

The Forecast Hub team used the submitted forecasts to generate three different ensemble forecasts. In the KITCOVIDhub-median_ensemble, the α-quantile of the ensemble forecast is obtained as the median of the α-quantiles of the member forecasts. In the KITCOVIDhub-mean_ensemble the mean instead of the median is applied for aggregation. In KITCOVIDhub-inverse_wis_ensemble, a convex combination of the α-quantiles of the member forecasts is used. 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 (Bracher et al., 2020) 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 (Reich et al., 2019a). This approach is referred to as Vincentization (in reference to Vincent (1912), see e.g., Busetti (2017)). 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 (2021) and Ray et al. (2022). Notably, Taylor and Taylor (2021) 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. (2022).

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.

4.3. Results

Figures 4.2 and 4.3 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). In each Figure, case and death forecasts for Germany are shown in panels (a) and (b), while the same for Poland is displayed in panels (c) and (d). 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 Figures C.1 and C.2 in the Appendix, and forecasts at horizons of three and four weeks are shown in Figures C.3–C.6 in the Appendix. 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.

4.3.1. Formal evaluation, January–April 2021

Table 4.2 and Figure 4.4 (panels (a), (b) for Germany and (c), (d) for Poland) 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 Table C.1 in the Appendix. The WIS can moreover be decomposed into components representing underprediction, forecast spread and overprediction (see Methods), which we show in Figure C.7. Detailed results in tabular form at horizons of three and four weeks ahead can be found in Table C.2. As specified in the study protocol, we also provide results for cumulative cases and deaths (Tables C.4 and C.5) and based on JHU rather than RKI/MZ data (Tables C.6 and C.7; 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 Figure C.8 in the Appendix.

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 4.4 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

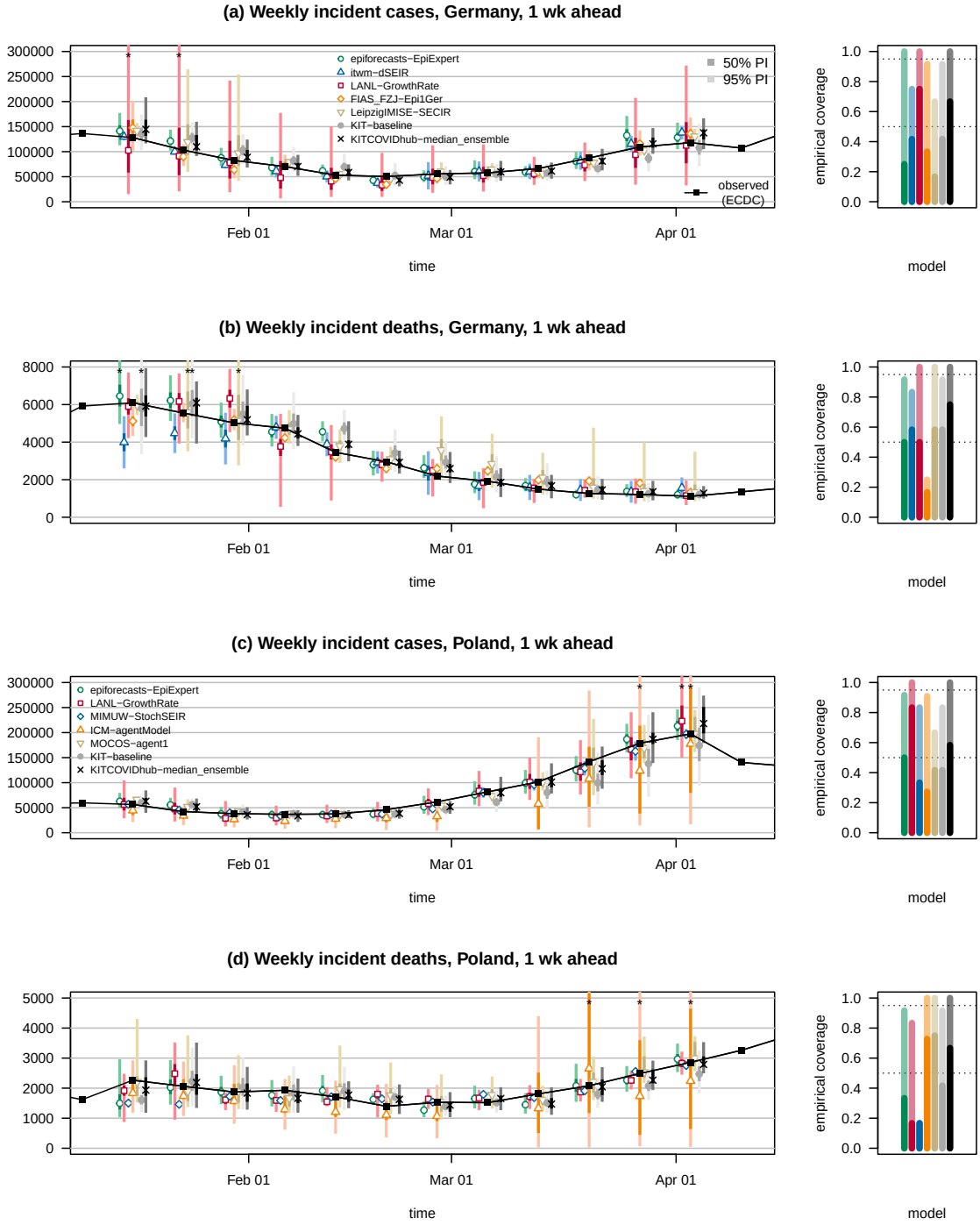


Figure 4.2.: One-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. The figure shows forecasts from a baseline model, the median ensemble of all submissions, and a subset of submitted models with above-average performance. The black line shows observed data. Colored 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 Figure C.1 in the Appendix. 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.

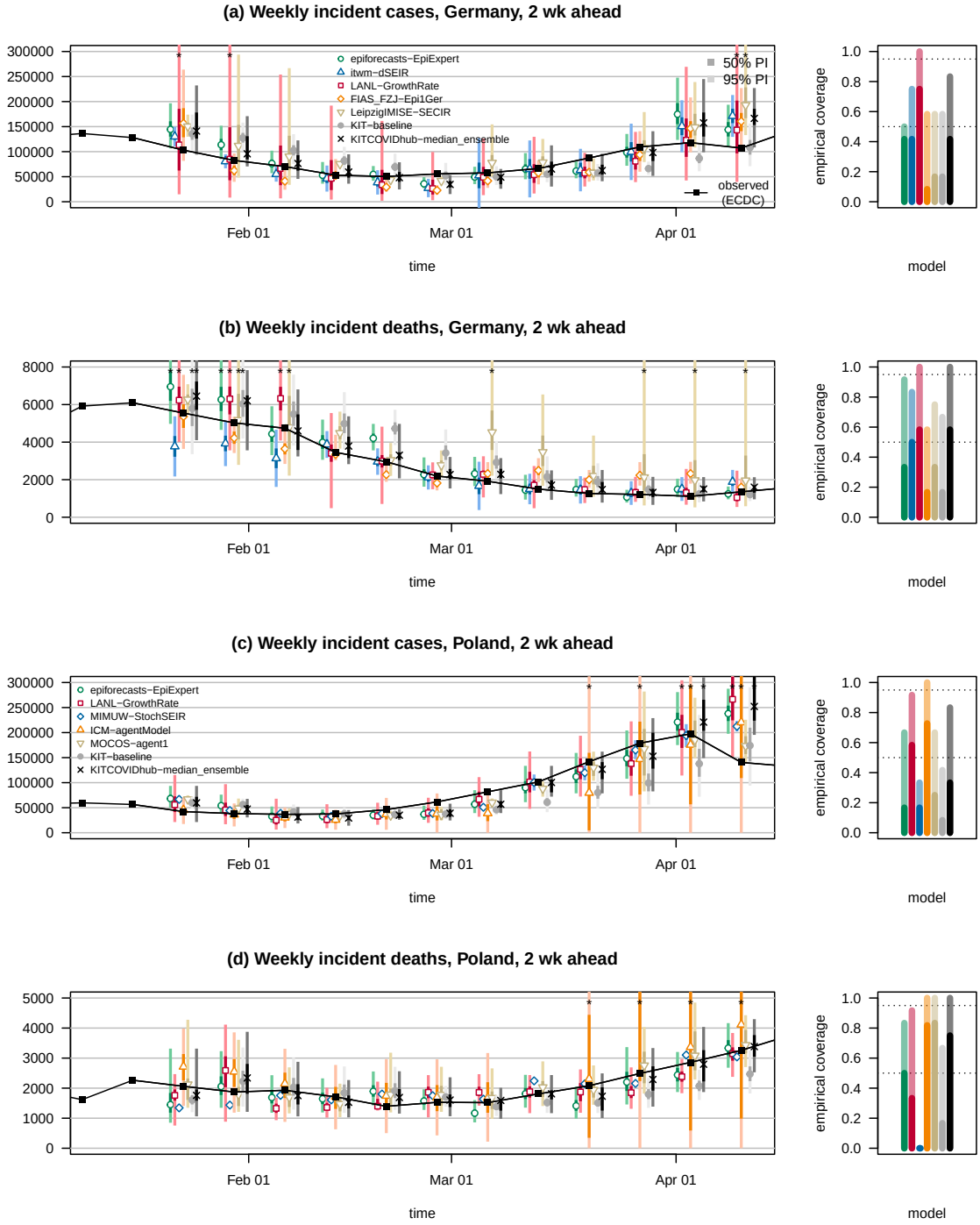


Figure 4.3.: Two-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. The figure shows forecasts from a baseline model, the median ensemble of all submissions and a subset of submitted models. The remaining submitted models are displayed in Figure C.2 in the Appendix. 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 4.2 for a detailed explanation of plot elements.

performance, does not yield any clear benefits over the unweighted median and mean ensembles. As can be seen from Figures C.9 and C.10 in the Appendix, 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 4.2 and 4.3, and resulted in large differences in the empirical coverage rates of 50% and 95% prediction intervals (Table 4.2 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 Figure C.7 in the Appendix (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).

Table 4.2.: Forecast evaluation for Germany and Poland (incidence scale, based on RKI/MZ data). Summaries are based on 12 weekly forecasts per target.

	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

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.

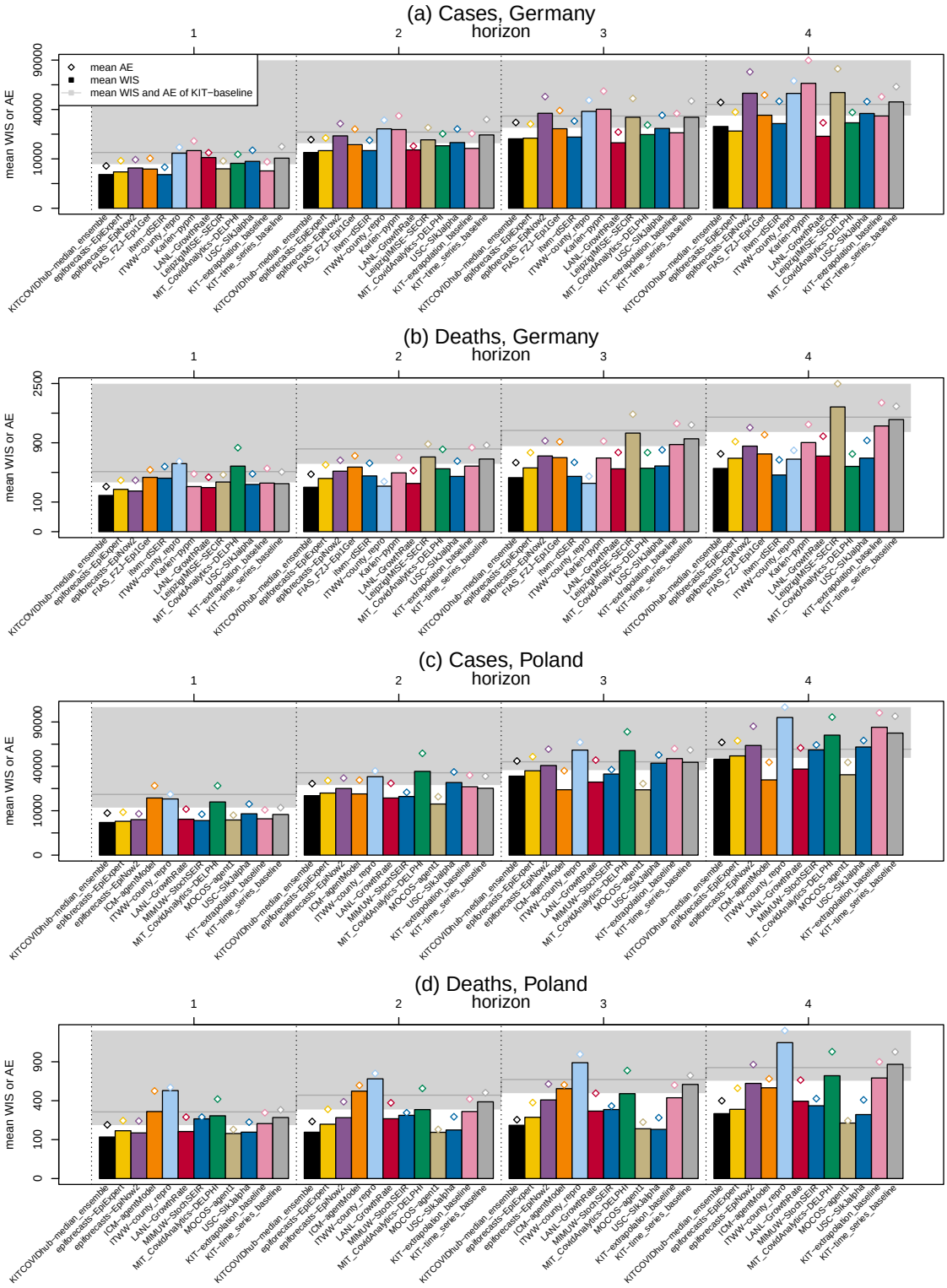


Figure 4.4.: Average weighted interval scores (bars) and absolute errors (diamonds) achieved by models across countries, targets and forecast horizons (12 weekly forecasts). The bottom end of the grey area represents the mean WIS of the baseline model KIT-baseline and the grey horizontal line its mean absolute error. Values are shown on a square-root scale to enhance readability. Only models covering all four horizons are shown.

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 4.3 provides a summary of the respective results at the one and two week horizons (results for three and four weeks can be found in Table C.3 in the Appendix). 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 4.2 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 (Bracher et al., 2020), 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 Tables C.8 and C.9 in the Appendix. 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

Table 4.3.: Forecast evaluation at the regional level, Germany and Poland (incidence scale, RKI/MZ data). Results are averaged over the different regions (16 states in Germany, 16 voivodeships in Poland).

	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

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.

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. (2021b).

4.3.2. 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.

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 4.1, 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 (Davies et al., 2021) 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 (Berndt et al., 2021a) 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 (Fischer-Fels, 2021) and the state of Baden-Württemberg (Landesgesundheitsamt Baden Württemberg, 2021), 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 (Robert Koch Institute, 2021b) 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

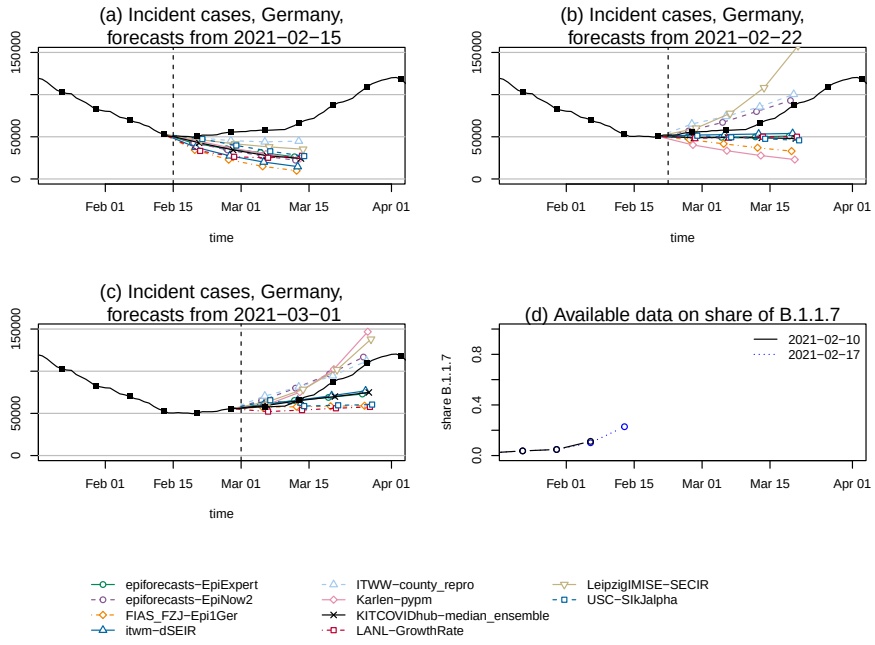


Figure 4.5.: Case forecasts in Germany preceding the upward trend change in March 2022. Point forecasts of cases in Germany, as issued on (a) 15 February, (b) 22 February and (c) 1 March 2021. These dates, shown as vertical dashed lines, mark the start of a renewed increase in overall case counts due 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.

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 4.5 shows the case forecasts of all submitted models and the median ensemble from (a) 15 February, (b) 22 February and (c) 1 March 2021. In panel (d) we moreover 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 Figure C.11 in the Appendix). 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 4.5). The other model including 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 (GISAID Initiative, 2021) only contained 2271 sequenced samples for Poland by 29 March 2021 (MI2 Data Lab, Warsaw University of Technology, 2021). 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 Figure C.1 in the Appendix).

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. Figure 4.6 shows forecasts made on (a) 22 March, (b) 29 March and (c) 5 April. It can be seen that the trajectory predicted by the two mentioned models differed substantially from those of most others, 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 4.2). In retrospective discussions, the respective teams noted

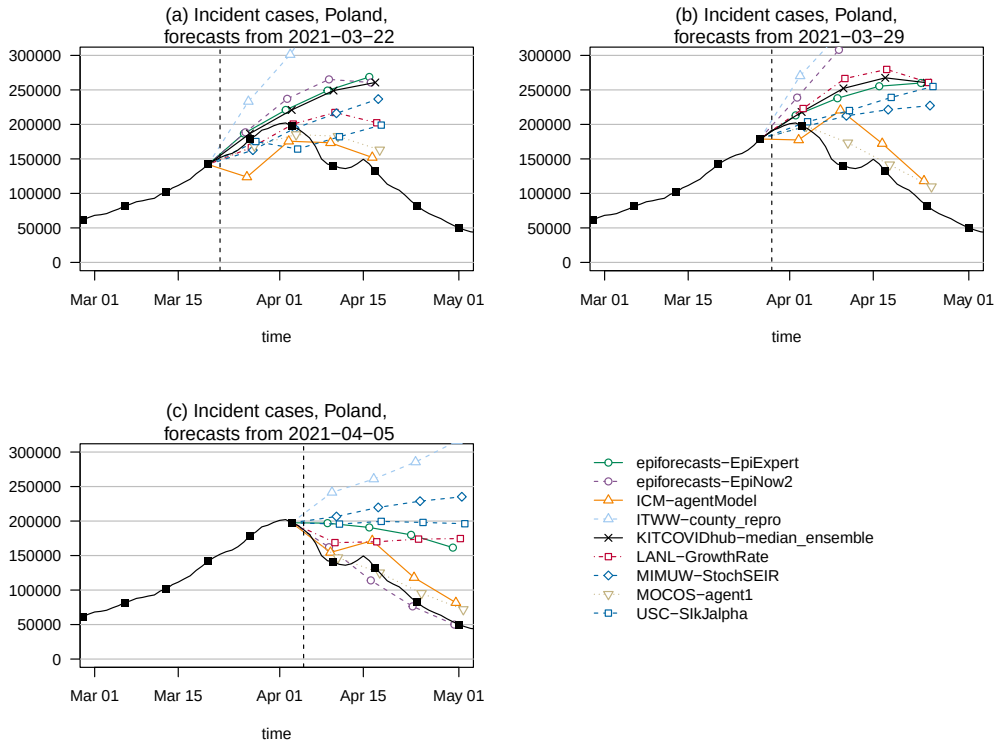


Figure 4.6.: Case forecasts in Poland surrounding the peak in April 2022. Point forecasts of cases in Poland from (a) 22 March, (b) 29 March and (c) 5 April 2021, surrounding the peak week. In each panel, the date at which forecasts were created is marked by a dashed vertical line. The models ICM-agentModel and MOCOC-agent1 anticipated the trend change correctly, while the remaining models show more or less pronounced overshoot.

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.

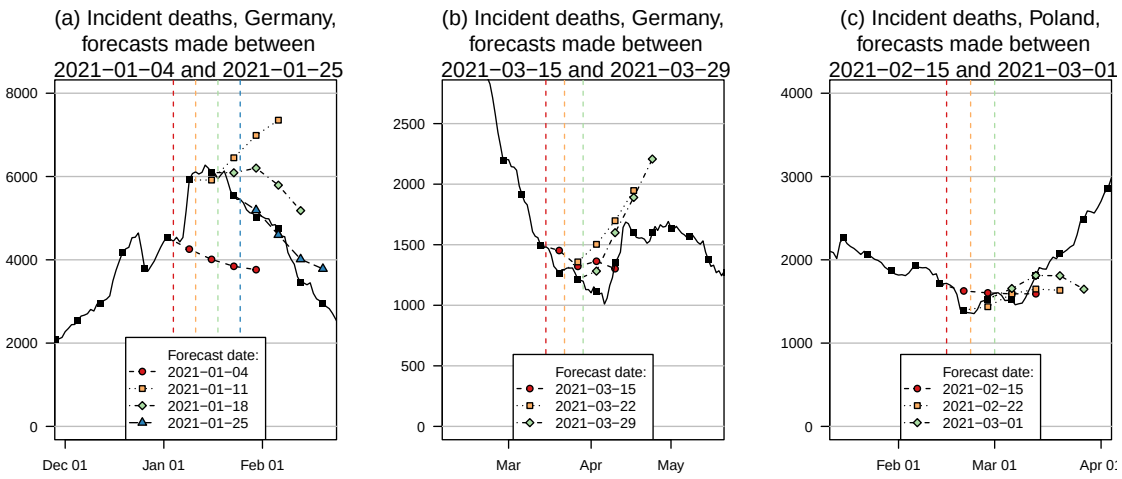


Figure 4.7.: Death forecasts preceding trend changes. Point forecasts of the median ensemble during changing trends in deaths. Panel (a): Downward turn in Germany, January 2021. Panel (b): Upward turn in Germany, March 2021. Panel (c): Upward turn in Poland, February/March 2021. Different colours and point/line shapes represent forecasts made at distinct time points (marked by dashed vertical lines).

In Germany, the study period coincided almost perfectly with a prolonged period of decline in deaths. In Figure 4.7, panels (a) and (b) show the behaviour of the median ensemble at 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 4.7, 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 4.2 and 4.3).

4.4. 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 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 (Bracher et al., 2021b) that meaningful case forecasts are only feasible at very short horizons. It also agrees with recent results from the US COVID-19 Forecast Hub (Reich et al., 2021), 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 (Keyel and Kilpatrick, 2021), 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 (Funk et al., 2020) and the US (Ray et al., 2021). To interpret these insights we note that, in principle, there are two ways of forecasting epidemiological time series. The first approach is to apply 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. The second approach is to establish 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 second category, 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 the first approach, which largely reduces to trend extrapolation, unless models are carefully tuned to changing NPIs (see Table 4.1). Theoretical arguments on the limited predictability of turning points in such curves have been brought forward (Castro et al., 2020; Wilke and Bergstrom, 2020), 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 (Harvey, 2021) 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 (McDonald et al., 2021). 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 (Arık et al., 2021; Fuhrmann and Barbarossa, 2020). 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 (Berndt et al., 2021b)). 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 (Lerch et al., 2017) in the literature and likewise occurs in, e.g., economics and meteorology (see illustrations in Table 1 from Lerch et al. (2017)). 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 (Coibion and Gorodnichenko, 2012).

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 by Taylor and Taylor (2021) and Ray et al. (2021) 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 (Ray et al., 2021) 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 (Sherratt et al., 2023), 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 case and death data used for evaluation. These have been taken from public sources of routine surveillance data (Robert Koch Institute, 2023; Polish Ministry of Health, 2022; Johns Hopkins University Center for Systems Science and Engineering, 2022), from which they can likewise be obtained. Forecasts can be visualised interactively at https://kitmetricslab.github.io/forecasthub/.

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 (Bracher et al., 2022). 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.

Appendix C

C.1. 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.

C.2. Additional forecast visualizations

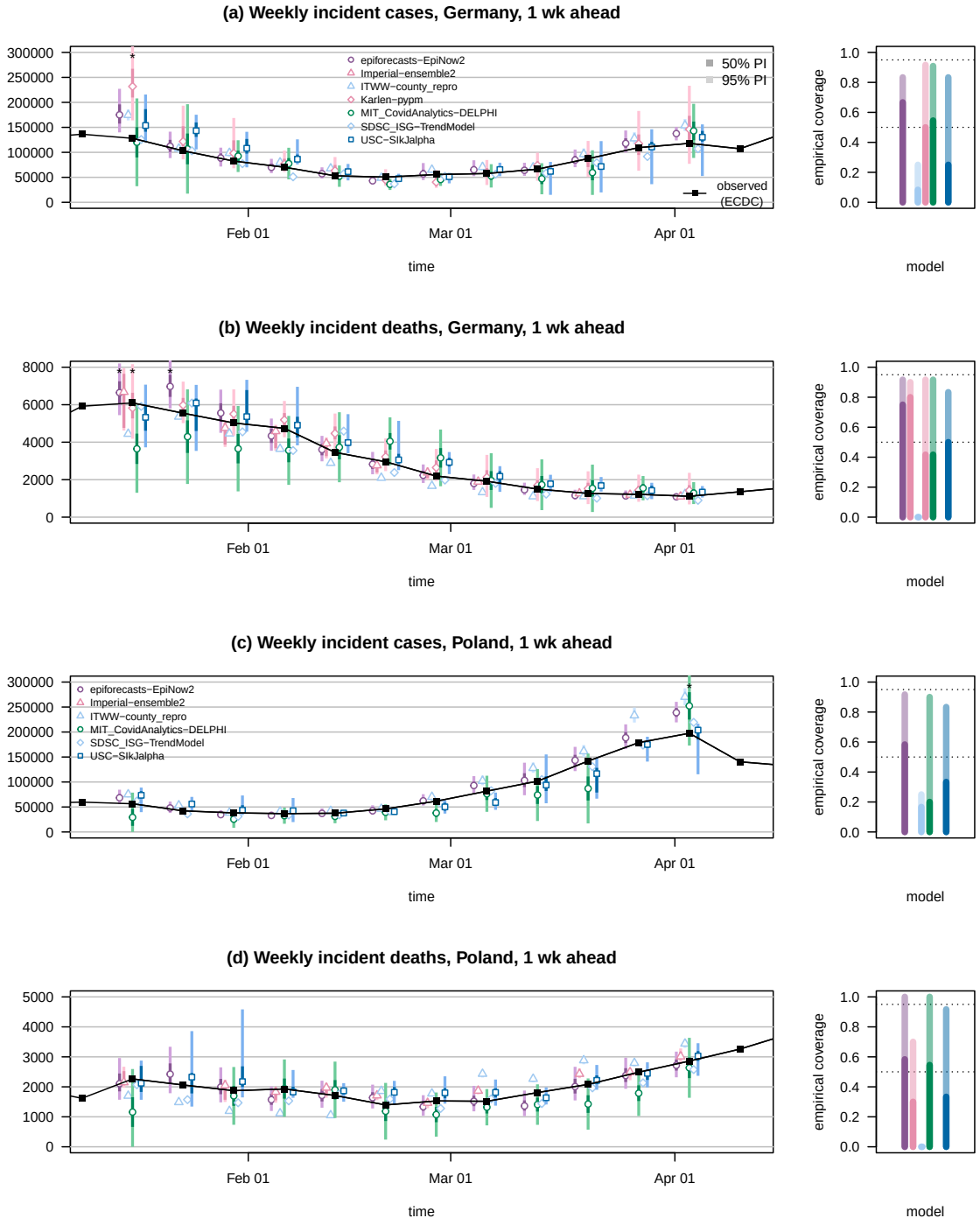


Figure C.1.: One-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. The figure shows forecasts from models not displayed in Figure 4.2. Asterisks mark prediction intervals exceeding the upper plot limit.

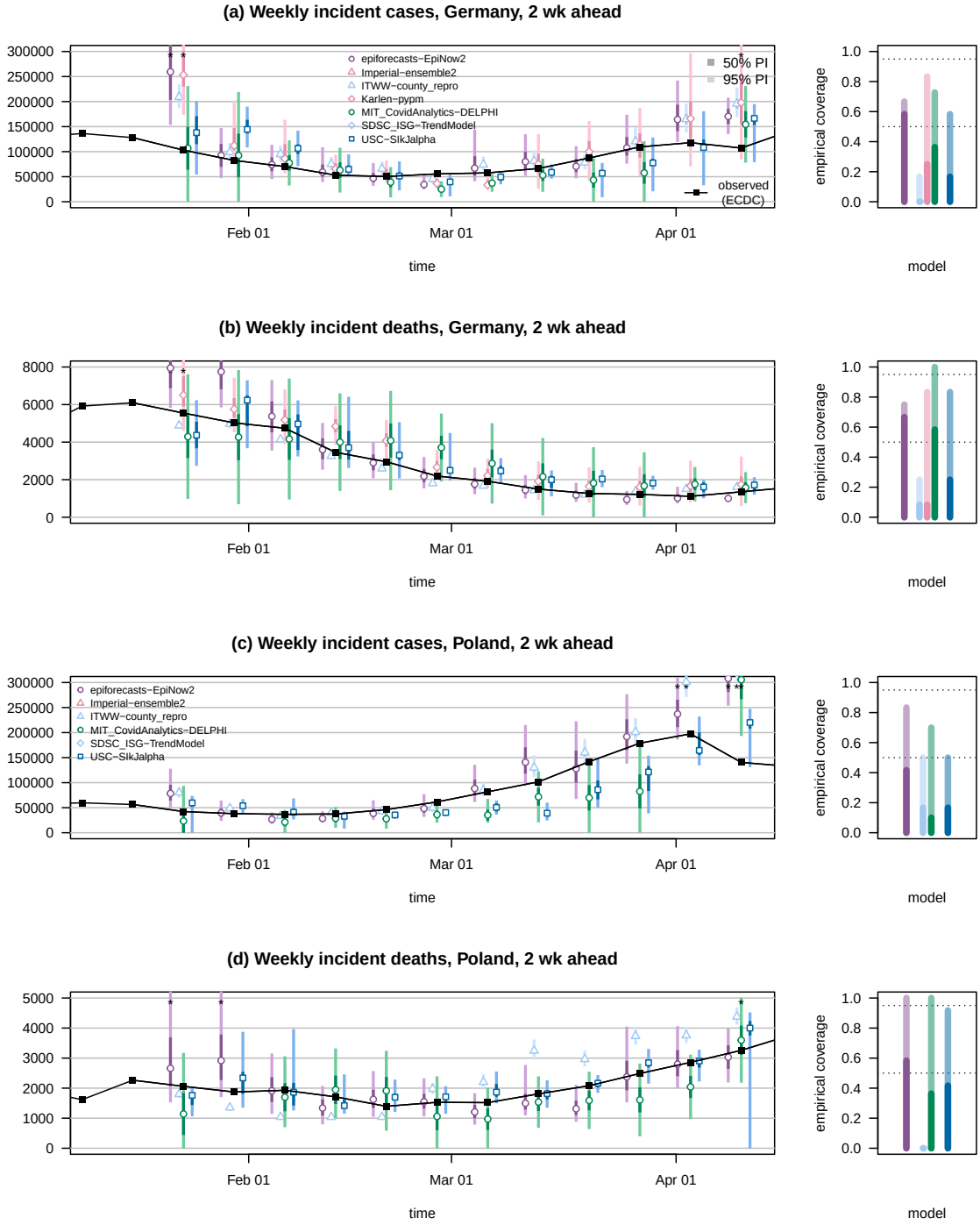


Figure C.2.: Two-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. The figure shows forecasts from models not displayed in Figure 4.3. Asterisks mark prediction intervals exceeding the upper plot limit.

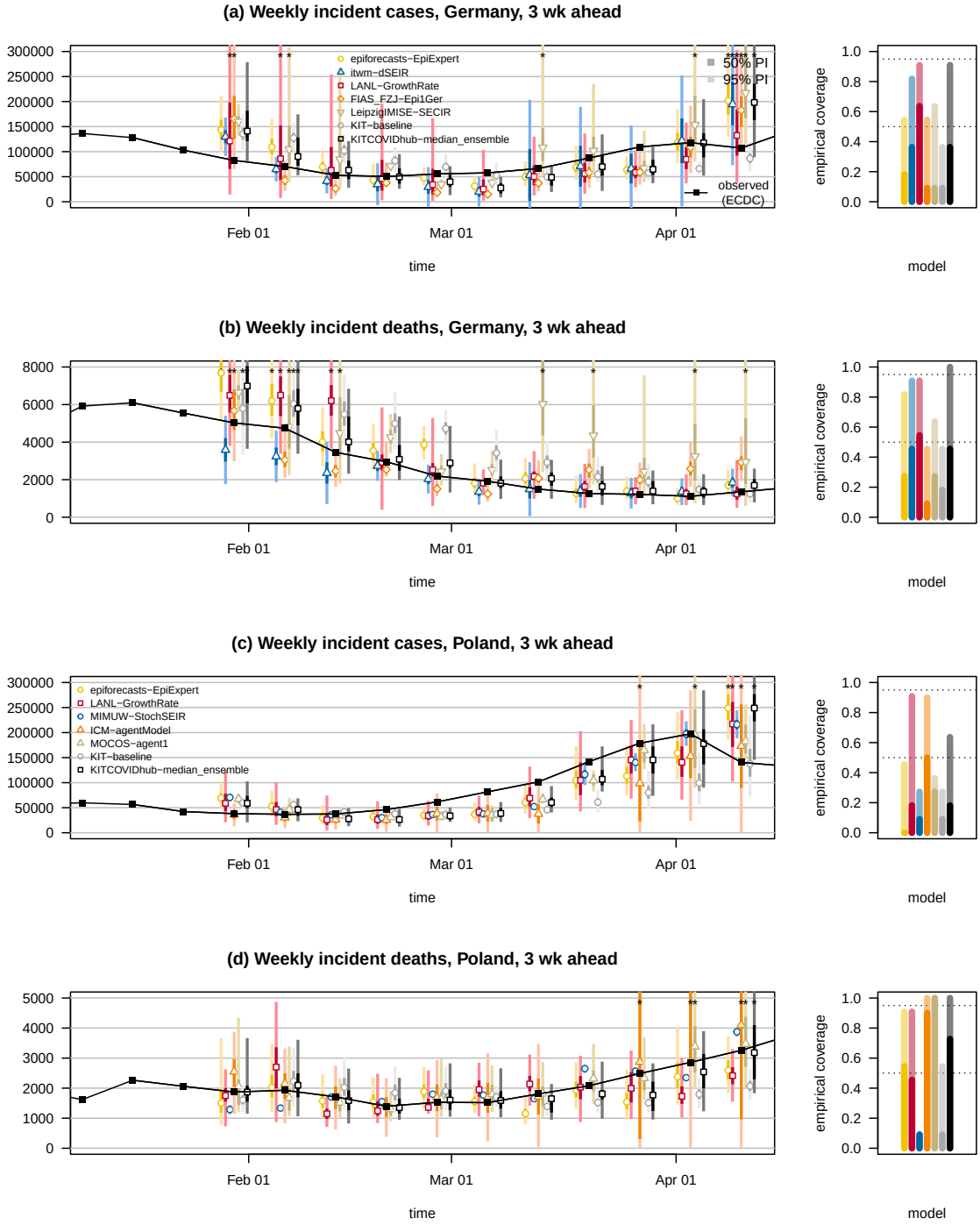


Figure C.3.: Three-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 4.2. Asterisks mark prediction intervals exceeding the upper plot limit.

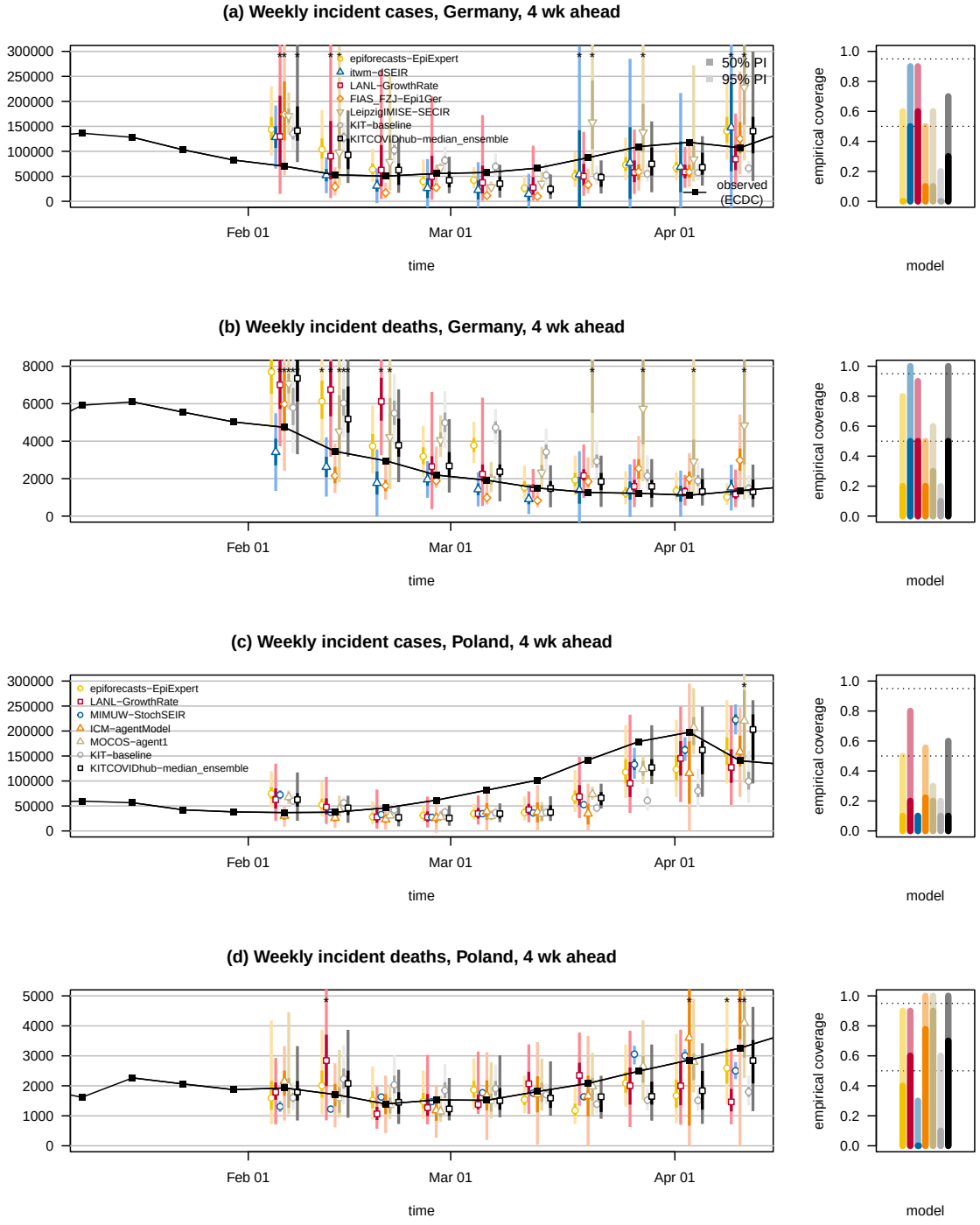


Figure C.4.: Four-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 4.2. Asterisks mark prediction intervals exceeding the upper plot limit.

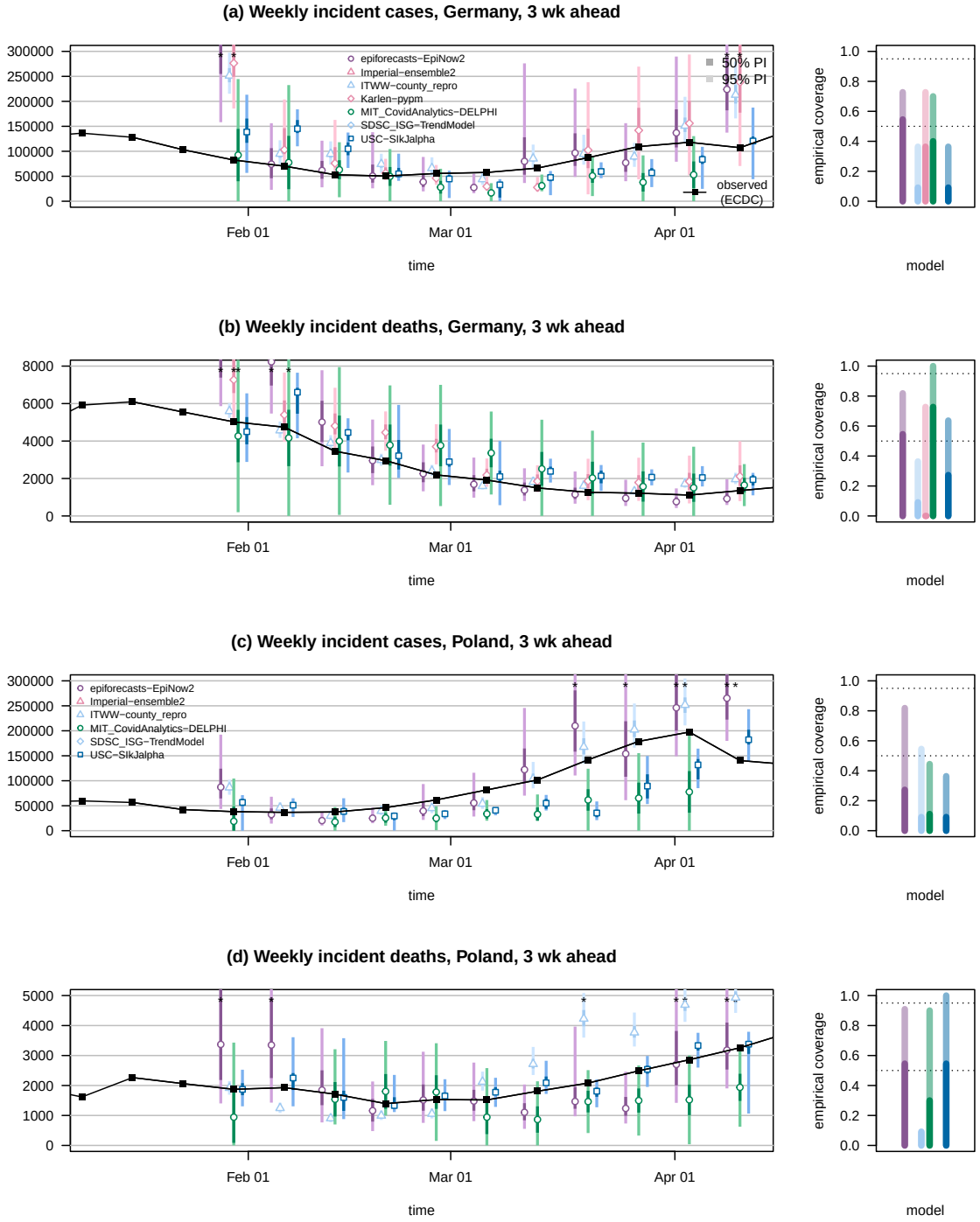


Figure C.5.: Three-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure C.1. Asterisks mark prediction intervals exceeding the upper plot limit.

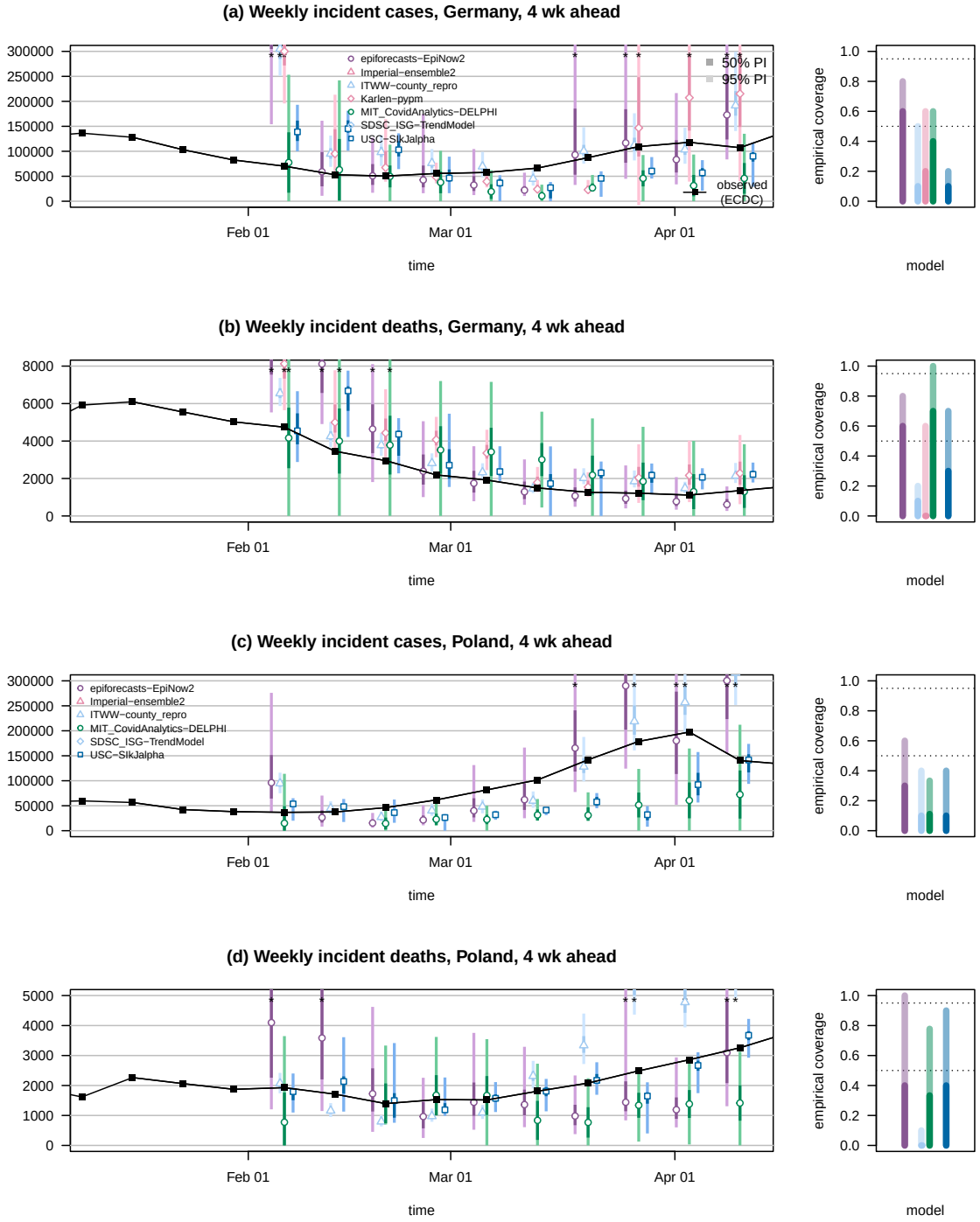


Figure C.6.: Four-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure C.1. Asterisks mark prediction intervals exceeding the upper plot limit.

C.3. Decomposition of average WIS values

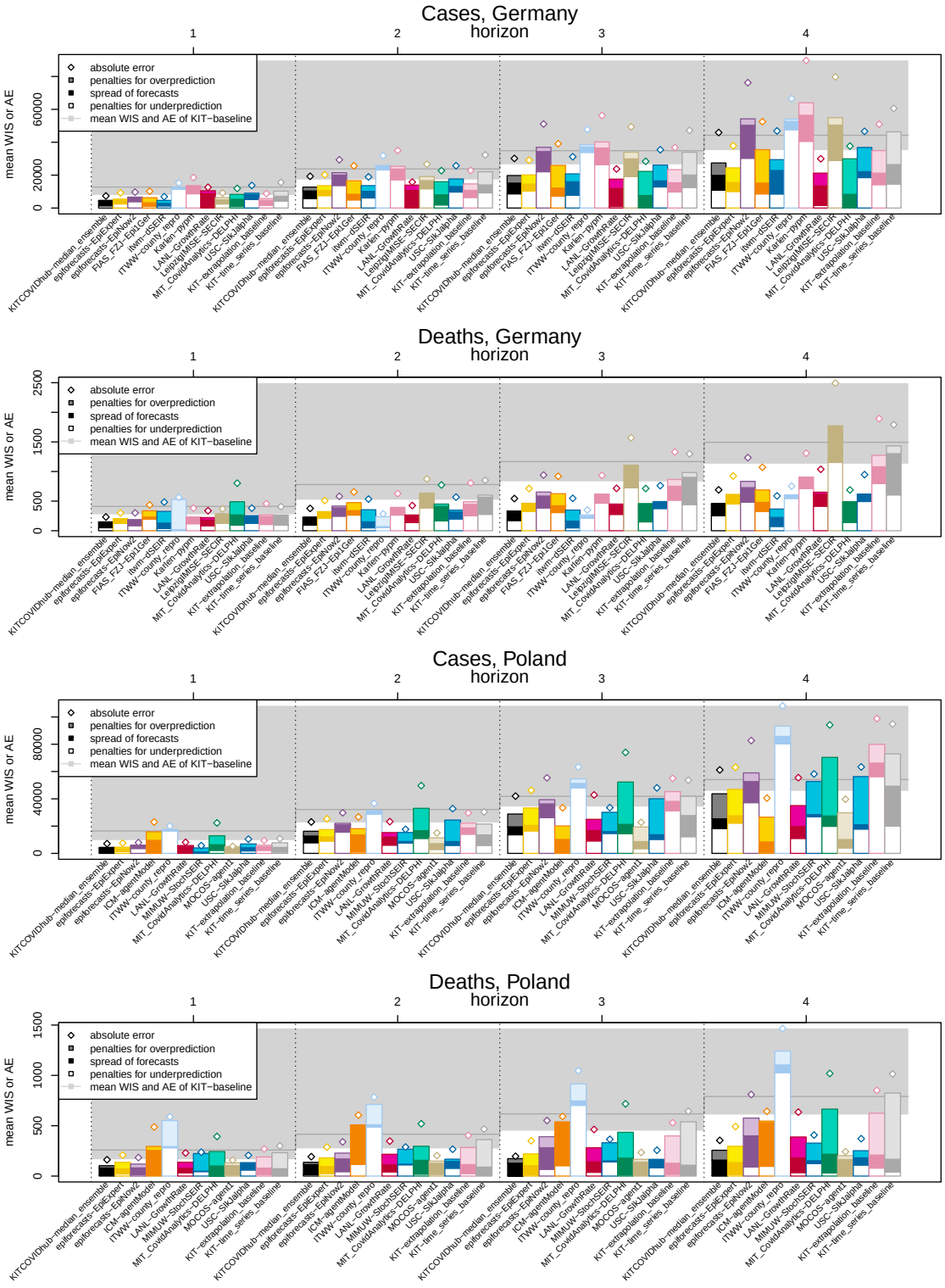


Figure C.7.: Average weighted interval score and absolute error achieved by models across countries, target,s and forecast horizons. The grey area represents the performance of the baseline model `KIT-baseline`. WIS values are decomposed into components for forecast spread, overprediction, and underprediction.

C.4. Individual WIS values

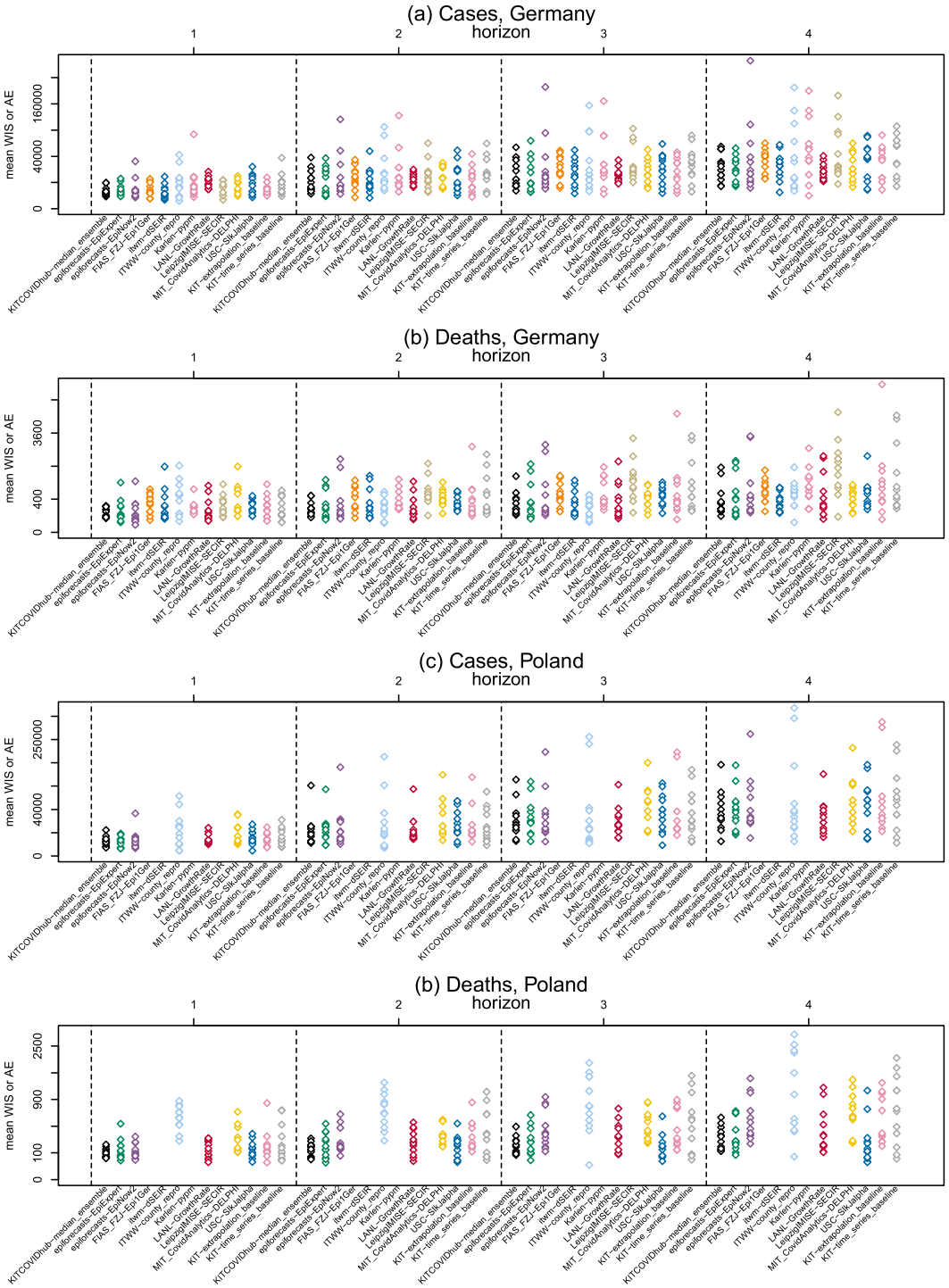


Figure C.8.: Individual weighted interval scores achieved by models across countries, targets, and forecast horizons. Each dot represents one score achieved over the 12-week evaluation period.

C.5. Weights in inverse-WIS ensembles

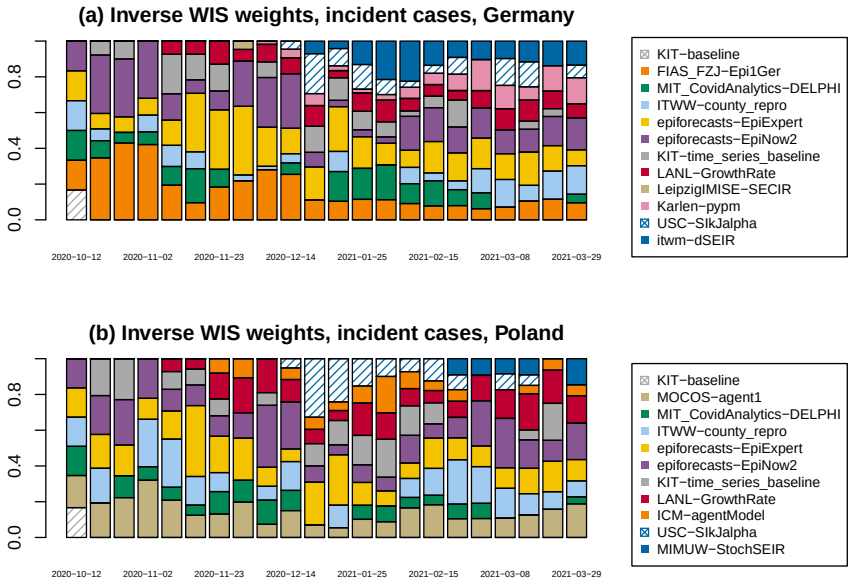


Figure C.9.: Weights in `KITCOVIDhub-inverse_wis_ensemble` for incident cases in Germany and Poland, October 2020–March 2021 (i.e., combined for the study periods of Bracher et al (2021) and the present manuscript).

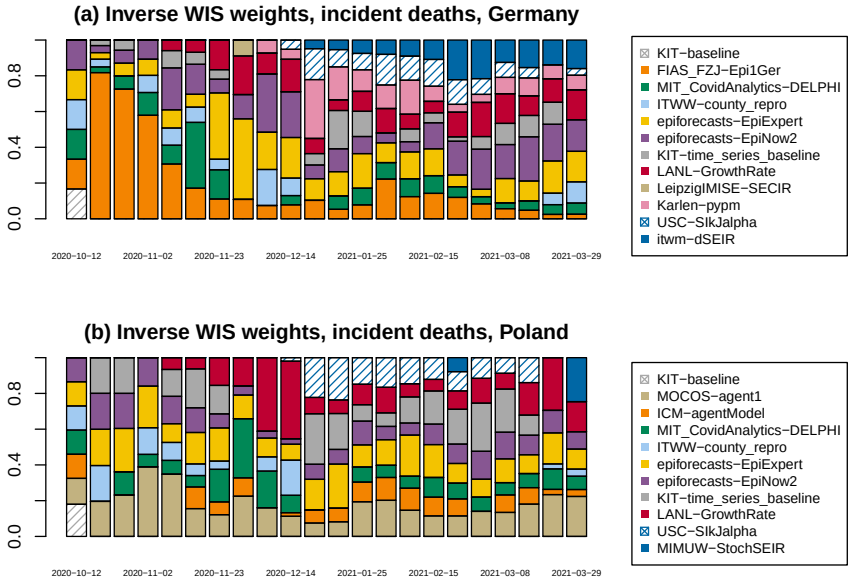


Figure C.10.: Weights in `KITCOVIDhub-inverse_wis_ensemble` for incident deaths in Germany and Poland, October 2020–March 2021 (i.e., combined for the study periods of Bracher et al (2021) and the present manuscript).

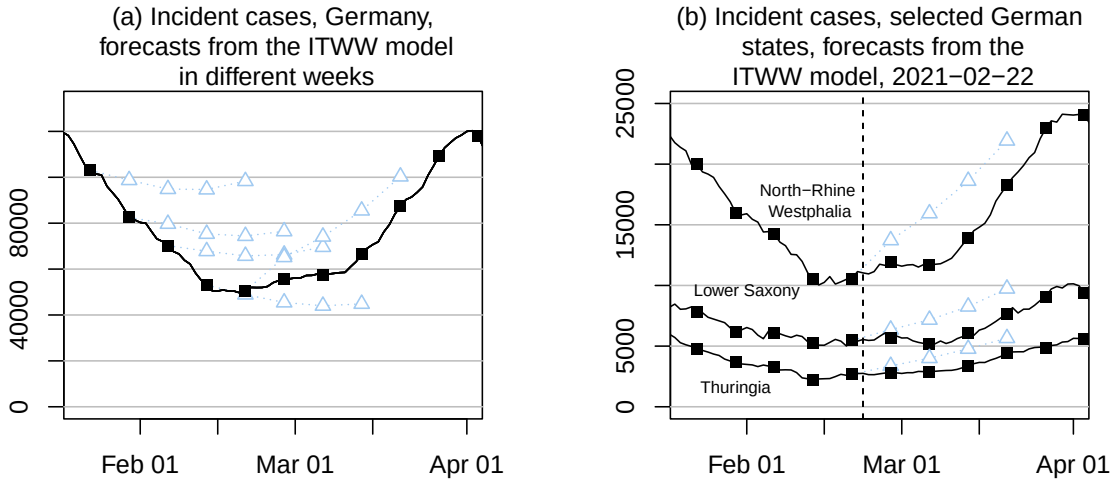


Figure C.11.: **a** Forecasts of cases in Germany by the `ITWW-county_repro` model, 25 January to 22 February 2021. **b** Forecasts for cases in selected German states by the `ITWW-county_repro` model, 22 February 2021.

C.6. Additional summary tables on forecast evaluation

In the following tables, 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 scriptsize of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

$C_{0.5}$ 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.

Table C.1.: 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 AE 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

Model	1 wk case		2 wk case		3 wk case		4 wk case		1 wk death		2 wk death		3 wk death		4 wk death	
	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

Model	1 wk case		2 wk case		3 wk case		4 wk case		1 wk death		2 wk death		3 wk death		4 wk death	
	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 C.2.: Forecast evaluation for Germany and Poland, 3 and 4 weeks ahead (incidence scale, based on RKI/MZ data).

Germany

Model	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
	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

Model	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
	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

Table C.3.: 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).

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

Table C.4.: Forecast evaluation for Germany and Poland, 1 and 2 weeks ahead (cumulative scale, based on RKI/MZ data).

Germany

Model	1 wk ahead cumul case				2 wk ahead cumul case				1 wk ahead cumul death				2 wk ahead cumul death			
	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

Model	1 wk ahead cumul case				2 wk ahead cumul case				1 wk ahead cumul death				2 wk ahead cumul death			
	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

Table C.5.: Forecast evaluation for Germany and Poland, 3 and 4 weeks ahead (cumulative scale, based on RKI/MZ data).

Germany

Model	3 wk ahead cumul case				4 wk ahead cumul case				3 wk ahead cumul death				4 wk ahead cumul death			
	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

Model	3 wk ahead cumul case				4 wk ahead cumul case				3 wk ahead cumul death				4 wk ahead cumul death			
	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

Table C.6.: Forecast evaluation for Germany and Poland, 1 and 2 weeks ahead (incidence scale, based on JHU data).

Germany

Model	1 wk ahead case (JHU)				2 wk ahead case (JHU)				1 wk ahead death (JHU)				2 wk ahead death (JHU)			
	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

Model	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
	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

Table C.7.: Forecast evaluation for Germany and Poland, 3 and 4 weeks ahead (incidence scale, based on JHU data).

Germany

Model	3 wk ahead case (JHU)				4 wk ahead case (JHU)				3 wk ahead death (JHU)				4 wk ahead death (JHU)			
	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

Model	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
	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

Table C.8.: Forecast evaluation for Germany and Poland, pooled across evaluation periods, 1 and 2 weeks ahead (incidence scale, based on RKI/MZ data).

Germany

Model	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
	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

Model	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
	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

Table C.9.: Forecast evaluation for Germany and Poland, pooled across evaluation periods, 3 and 4 weeks ahead (incidence scale, based on RKI/MZ data).

Germany

Model	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
	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			0.33	0.83			0.42	0.92			0.42	0.92			0.42	1.00
itwm-dSEIR	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
ITWW-county_repro			0.33	0.75			0.17	0.67			0.00	0.75			0.00	0.67
Karlen-pypm	*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
LANL-GrowthRate	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
LeipzigIMISE-SECIR	*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
MIT_CovidAnalytics-DELPHI																
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

Model	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
	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

5. Model diagnostics and forecast evaluation for quantiles

5.1. Background and motivation

Quantiles play ubiquitous roles in statistical modeling. Univariate probability distributions are characterized by their quantiles. So, in a sense, if we are concerned with the estimation and prediction of individual variables, distributional approaches can be thought of as quantile models. Quantile regression has become one of the most fundamental and widely-used tools for statistical analysis (Koenker and Bassett, 1978; Koenker, 2005, 2017) and links to robustness (Huber and Ronchetti, 2009). In the machine learning community, interest in quantile-based methods has been strongly on the rise (Gasthaus et al., 2019; Chung et al., 2021). Quantiles at specific levels occur as decision thresholds and optimal point forecasts in a plethora of practically relevant situations, ranging from Value-at-Risk in financial regulation (Basle Committee on Banking Supervision, 1996; Duffie and Pan, 1997) and Growth-at-Risk (Adrian et al., 2019) to renewable energy markets and the classical newsvendor problem (Pinson et al., 2007; Gneiting, 2011b). Frequently, confidence intervals and interval forecasts are also defined via quantiles. For example, the widely used equal-tailed form of the 90% interval ranges from the predictive quantile at level 0.05 to that at level 0.95, an upper 90% interval is set via the quantile at level 0.1, and a lower 90% interval via the quantile at level 0.9.

Model diagnostics and forecast evaluation are crucial and closely related tasks in these settings. While model diagnostics seeks to assess and quantify in-sample goodness (or lack) of fit, forecast evaluation is concerned with predictive performance out-of-sample. In this paper, we review tools for model checking and the evaluation of predictive performance when forecasts are cast in the form of quantiles or quantile-bounded intervals. We

leverage a recently proposed, comprehensive theory of calibration (Gneiting and Resin, 2023), which covers all types of statistical functionals induced by identification functions. We apply this theory to the case of quantiles, and illustrate the respective diagnostic and inferential tools on key real-world examples covering a wide range of application fields such as epidemiology, economics, and energy. A common theme is that we take discrete data at face value, and review methods that explicitly take account thereof. The discreteness seen in many real-world datasets poses technical challenges for quantile-based models, as nominal levels generally cannot be met exactly, and the treatment of these issues has seen limited attention in the literature, with notable exceptions, such as the work of Gelman et al. (2000), Czado et al. (2009) and Homburg et al. (2019).

To illustrate the tools discussed in this paper, we use data from the United States (US) COVID-19 Forecast Hub (Cramer et al., 2022a,b) as running example. The Forecast Hub is a public platform assembling forecasts of confirmed cases, hospitalizations and deaths from COVID-19 in the US from more than 80 different models run by academic, industry and independent research groups. The forecasts have been collected in a weekly rhythm since April 2020 in the form of 23 predictive quantiles at levels 0.01, 0.025, 0.05, 0.1, ..., 0.95, 0.975, and 0.99. Forecasts are issued at the state, national, and for certain targets, county levels and refer to data from the Johns Hopkins Center for Systems Science and Engineering. In this paper, we focus on state- and national-level death predictions at a lead time of one week ahead. Figure 5.1 illustrates forecasts from three different models at the national level as issued in April through December 2021. COVIDhub-baseline is a simplistic reference model that always uses the last available observation as its predictive median; the remaining quantiles are estimated based on past forecast errors (Cramer et al., 2022b). COVIDhub-ensemble is the main output of the platform and shown by default in its dashboard. Its definition has changed several times; it was a quantile-wise mean of all submitted forecasts until July 2020, when this was replaced by a quantile-wise median. Starting from November 2021, a scheme has been in use which involves the selection and weighting of ten member models based on recent performance (Ray et al., 2022). The KITmetricslab-select_ensemble follows a similar principle and is based on a data-driven forward selection and linear combination of best performing models in the recent past.

The remainder of the paper is organized as follows. Section 5.2 is concerned with notions of calibration for quantiles. In a nutshell, the term calibration or reliability refers to the statistical consistency between the modeled or predicted quantiles and the outcomes. We distinguish unconditional calibration, which corresponds to classical coverage criteria, with caveats in discrete cases, from the stronger notion of conditional calibration, as can be visualized in quantile reliability diagrams.

Section 5.3 turns to consistent loss or scoring functions for quantiles — including, but not limited to, the popular asymmetric piecewise linear score or pinball loss — that serve dual uses as loss functions in estimation problems, and for comparative assessment and ranking in prediction problems. Briefly, a scoring function is consistent for the α-quantile if the expected loss under an outcome with distribution F is minimized by issuing the α-quantile $q_\alpha(F)$ as point forecast, and the pinball loss is the most prominent function of this type. We discuss decompositions of mean scores into miscalibration, discrimination, and uncertainty components, and link to the coefficient of determination and skill scores. Murphy diagrams leverage mixture representations for scoring functions and allow researchers to assess performance in terms of just any consistent scoring function simultaneously.

In Section 5.4, we illustrate the use of these diagnostic and inferential tools on Engel's food expenditure data, where we contrast in-sample model diagnostics and out-of-sample forecast evaluation, and the Global Energy Forecasting Competition 2014.

The paper closes with a discussion in Section 5.5. Importantly, the methods and tools we discuss adhere to the prequential principle (Dawid, 1984), as they are based on pairs of the (quantile) forecast and the respective outcome only, without considering the generating methods or mechanism, and they apply in full generality, without dependence on data structures. They are relevant whenever a collection of modeling or forecast cases is to be assessed, be it in unstructured, time series, spatial, spatio-temporal, or just any type of setting. Software for the application of these tools and replicating the results in the paper is available in R (R Core Team, 2021).

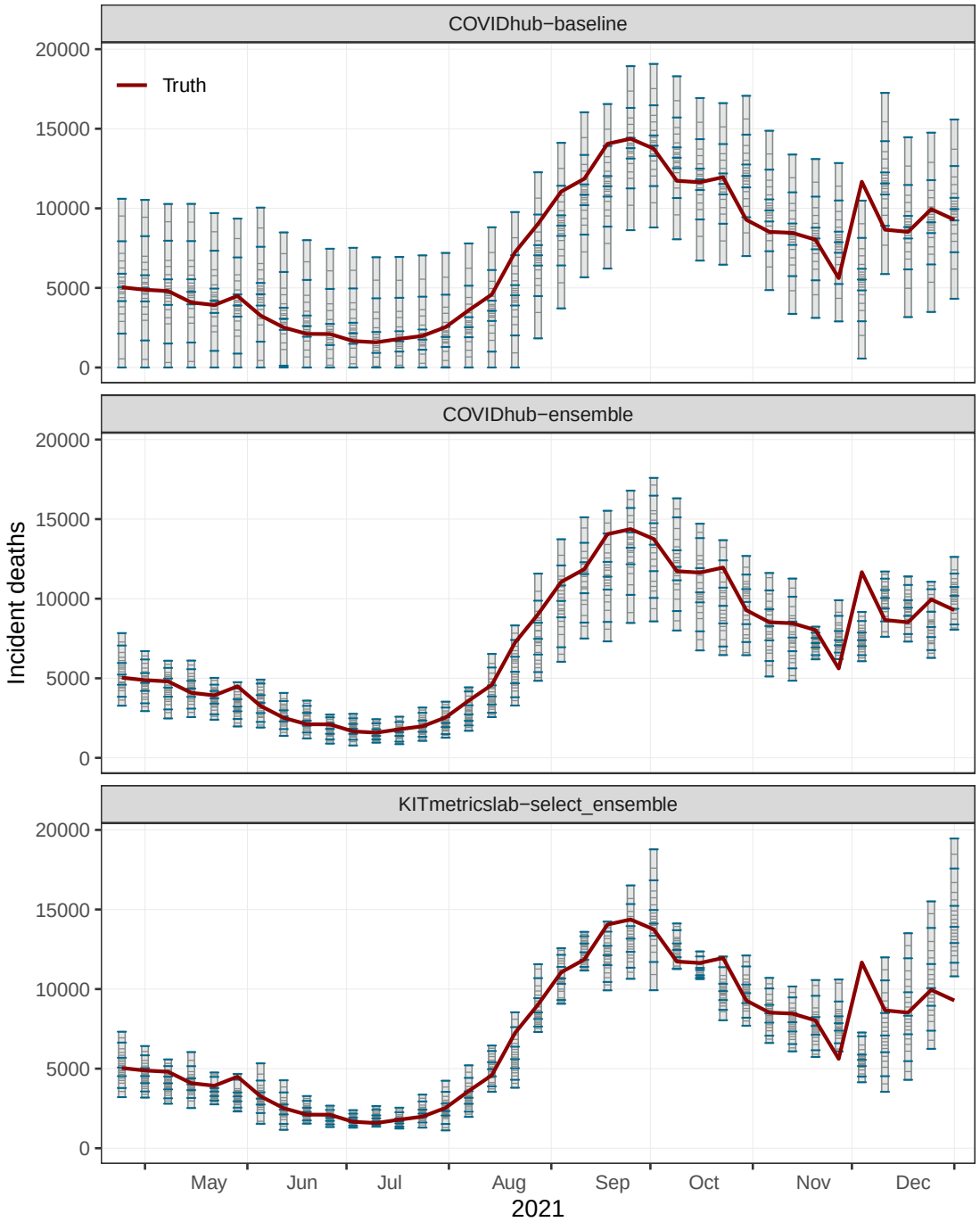


Figure 5.1.: Quantile forecasts of weekly deaths from COVID-19 in the US from the COVIDhub-baseline, COVIDhub-ensemble and KITmetrics-select_ensemble models. We show predictive quantiles valid for the 37 weeks starting April 24, 2021 at 23 levels, with those at levels 0.01, 0.05, 0.25, 0.5, 0.75, 0.95, and 0.99 highlighted in blue. The red lines show the observed counts.

5.2. Conditional and unconditional quantile calibration

For a probability distribution with strictly increasing cumulative distribution function (CDF) F, the α-quantile ($0 < \alpha < 1$) is the unique number x such that $F(x) = \alpha$. In the general setting of an increasing, but not necessarily strictly increasing CDF, we distinguish the lower α-quantile, $q_\alpha^-(F) = \sup\{x : F(x) < \alpha\}$, and the upper α-quantile, $q_\alpha^+(F) = \inf\{x : F(x) > \alpha\}$, and we refer to any number x that satisfies $q_\alpha^-(F) \leq x \leq q_\alpha^+(F)$ as an α-quantile of F. Equivalently, one may think of the α-quantile as a set-valued functional that assigns the closed interval

$$Q_\alpha(F) = \left[q_\alpha^-(F), q_\alpha^+(F)\right]$$

to the probability distribution F. The lower and upper quantiles need not coincide, and $Q_\alpha(F)$ is a non-degenerate interval if, and only if, the CDF F has a horizontal segment at the height α. Importantly, the α-quantile can be identified using the function

$$V_\alpha(x, y) = \mathbb{1}\{y < x\} - \alpha,$$

where the indicator function $\mathbb{1}\{\cdot\}$ equals 1 if the argument in brackets is true and 0 otherwise. Specifically,

$$q_\alpha^-(F) = \sup\left\{x : \int V_\alpha(x, y)\,\mathrm{d}F(y) < 0\right\}, \quad q_\alpha^+(F) = \inf\left\{x : \int V_\alpha(x, y)\,\mathrm{d}F(y) > 0\right\},$$

where the integral can be expressed equivalently as the expectation $\mathbb{E}_{Y\sim F}\left[V_\alpha(x, Y)\right]$. In this light, V_α is referred to as an identification function for the α-quantile. These relationships become critical in the theory of quantile calibration to which we turn now. Readers with interest in practical considerations only may wish to ignore the population setting and move ahead to Sections 5.2.2 and 5.2.3, where we discuss diagnostic tools such as coverage plots and quantile reliability diagrams.

5.2.1. Quantile calibration in the prediction space setting

We now describe notions of calibration in the population setting of Gneiting and Ranjan (2013) and Gneiting and Resin (2023), and discuss calibration based on the joint distri-

bution of a bivariate random vector (X, Y), where X is conceptualized as a single-valued forecast and Y as the associated outcome. Throughout the section, we denote the joint distribution of (X, Y) by $\mathbb{P}$, and expectations are with respect to this measure. Generally, we adopt the theory of Gneiting and Resin (2023), which develops criteria of calibration that are satisfied by ideal forecasts, where a forecast is ideal if it arises under measure theoretic conditioning on information sets. We adapt and apply this theory to single-valued point forecasts whose merits as predictive α-quantiles are to be assessed.

The most basic notion is unconditional α-quantile calibration. Specifically, X is unconditionally α-quantile calibrated for Y if

$$\mathbb{E}\left[V_\alpha(X-\varepsilon, Y)\right] \leq 0 \quad \text{and} \quad \mathbb{E}\left[V_\alpha(X+\varepsilon, Y)\right] \geq 0 \quad \text{for all} \quad \varepsilon > 0,$$

or, equivalently,

$$\mathbb{P}\left(Y < X\right) \leq \alpha \quad \text{and} \quad \mathbb{P}\left(Y \leq X\right) \geq \alpha. \tag{5.1}$$

Evidently, these are versions of the classical non-exceedance criterion for quantiles that take discreteness into account. If the bivariate distribution of (X, Y) is continuous, then $\mathbb{P}(Y = X) = 0$, and the conditions in Equation 5.1 reduce to the equation $\mathbb{P}(Y \leq X) = \alpha$. Analogous considerations apply to quantile-bounded prediction intervals.

We turn to a stronger notion of calibration, namely, conditional α-quantile calibration. In a nutshell, this requires that unconditional α-quantile calibration continues to hold if we stratify by forecast value, regardless of the forecast value that we condition on. Technically, X is conditionally α-quantile calibrated for Y if

$$\mathbb{E}\left[V_\alpha(X-\varepsilon, Y) \mid X\right] \leq 0 \quad \text{and} \quad \mathbb{E}\left[V_\alpha(X+\varepsilon, Y) \mid X\right] \geq 0 \quad \text{for all} \quad \varepsilon > 0$$

almost surely or, equivalently,

$$\mathbb{P}\left(Y < X \mid X\right) \leq \alpha \quad \text{and} \quad \mathbb{P}\left(Y \leq X \mid X\right) \geq \alpha \quad \text{almost surely.} \tag{5.2}$$

If the bivariate distribution of (X, Y) is continuous, these conditions reduce to the equation $\mathbb{P}\left(Y \leq X \mid X\right) = \alpha$ (Krüger and Ziegel, 2021, Definition C.1). Clearly, conditional calibration implies unconditional calibration (Gneiting and Resin, 2023, Theorem 2.11).

Stated slightly informally, the α-quantile reliability diagram for X as a predictor of Y plots the graph of the function

$$x \mapsto q_\alpha\left(Y \mid X = x\right), \tag{5.3}$$

on the support of X, where for simplicity we assume that the α-quantile is a singleton.[1] In simple words, we plot the α-quantile of the conditional distribution of the outcome, given the prediction, versus the prediction. Evidently, X is α-quantile calibrated for Y if, and only if, the graph is (a subset of) the diagonal.

5.2.2. Unconditional calibration

We move on to empirical settings, in which we are given tuples of the form

$$(x_1, y_1), \ldots, (x_n, y_n), \tag{5.4}$$

where x_i is a posited α-quantile and y_i is the associated outcome, for $i = 1, \ldots, n$. A prediction x_i typically depends on a covariate or feature vector and represents a conditional quantile, but this is immaterial if we adhere to the prequential principle (Dawid, 1984), and so we leave the conditioning implicit. For example, the predictive quantiles may stem from a parametric quantile regression model (Koenker, 2005, 2017), a nonparametric statistical or machine learning model (Gasthaus et al., 2019; Henzi et al., 2021), or a contribution to a forecast competition.

Given such a collection of prediction-outcome pairs, we denote the lower and upper coverage by

$$\mathrm{c}_\alpha^- = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i < x_i\} \quad \text{and} \quad \mathrm{c}_\alpha^+ = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i \leq x_i\}, \tag{5.5}$$

respectively. If we apply the unconditional calibration condition from Equation 5.1 to the empirical measure of the data, we obtain the empirical unconditional calibration criterion

$$\mathrm{c}_\alpha^- \leq \alpha \leq \mathrm{c}_\alpha^+.$$

[1] For a rigorous, measure-theoretic treatment in the general setting of identifiable functionals see Section 2.4 in Gneiting and Resin (2023).

Therefore, to diagnose unconditional calibration it suffices to check whether the closed coverage interval $[\mathrm{c}_\alpha^-, \mathrm{c}_\alpha^+]$ contains the nominal level α.

To quantify sampling variability we consider consistency and confidence intervals. Consistency intervals describe how much coverage fluctuates around the nominal level α under the hypothesis of unconditional calibration and quantify how likely or unlikely an observed coverage is. In contrast, confidence intervals describe where the unknown (true) lower or upper coverage might lie, given the data at hand, and are positioned around c_α^- and c_α^+.

We construct consistency intervals from one-sided binomial tests with test statistics nc_α^- and nc_α^+ to check the conditions in Equation 5.1, corresponding to proper lower or proper upper coverage, respectively. The $(1-\beta) \times 100\%$ consistency interval with $\beta \in (0, 1)$ is framed by the critical values at a significance level of $\frac{\beta}{2}$ (Bonferroni correction), which derive from the binomial distribution $B(n, \alpha)$ as in the classical one-sided test. A combined test at level β suggests a lack of unconditional calibration if the coverage and consistency intervals fail to overlap. For the $(1-\beta) \times 100\%$ confidence intervals, we consider lower coverage and upper coverage separately. We resample observed prediction-outcome pairs to generate bootstrap values of the two types of coverage, and the confidence intervals are framed by the empirical quantiles of the bootstrap distributions at levels $\frac{\beta}{2}$ and $1 - \frac{\beta}{2}$.

If posited quantile forecasts at several levels are available, lower and upper coverage are conveniently shown in a coverage plot, as illustrated in Figure 5.2 and Figure 5.3 on forecasts at the 23 aforementioned levels from the US COVID-19 Forecast Hub. Both figures show consistency bands positioned around the diagonal (top row), and confidence bands (bottom row) positioned around the coverage at hand. Similar displays can be used to assess unconditional calibration of quantile-bounded interval forecasts.

At the national level as displayed in Figure 5.2, ties between quantile forecasts and observed counts hardly ever occur, so lower and upper coverage coincide. Figure 5.3 illustrates challenges with discrete data on forecasts for the state of Vermont, where there is a stark contrast between lower and upper coverage.

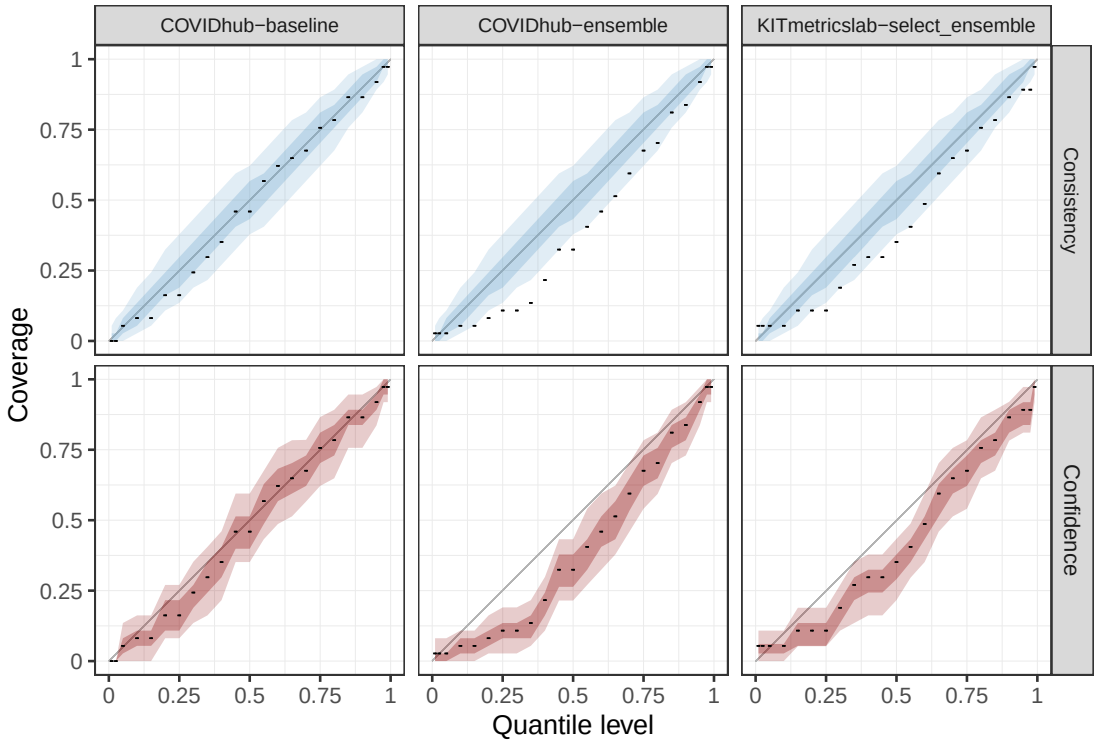


Figure 5.2.: Coverage plots for US COVID-19 Forecast Hub predictions at the national level, for which lower and upper coverage coincide. The colored bands show 50% and 90% consistency (top) and confidence intervals (bottom).

The importance of considering both lower and upper coverage in discrete settings is highlighted by the KITmetricslab-select_ensemble, which shows coverage well in line with the assumption of unconditional calibration. If we had ignored the discreteness and considered lower or upper coverage only, we would have misjudged forecast reliability.

Checks for unconditional calibration are critically important in the evaluation of predictive performance out-of-sample. For in-sample model diagnostics, they are largely irrelevant, as estimation via empirical score minimization frequently guarantees unconditional calibration, see Section 5.3.1 for theoretical results and Section 5.4.1 for illustration.

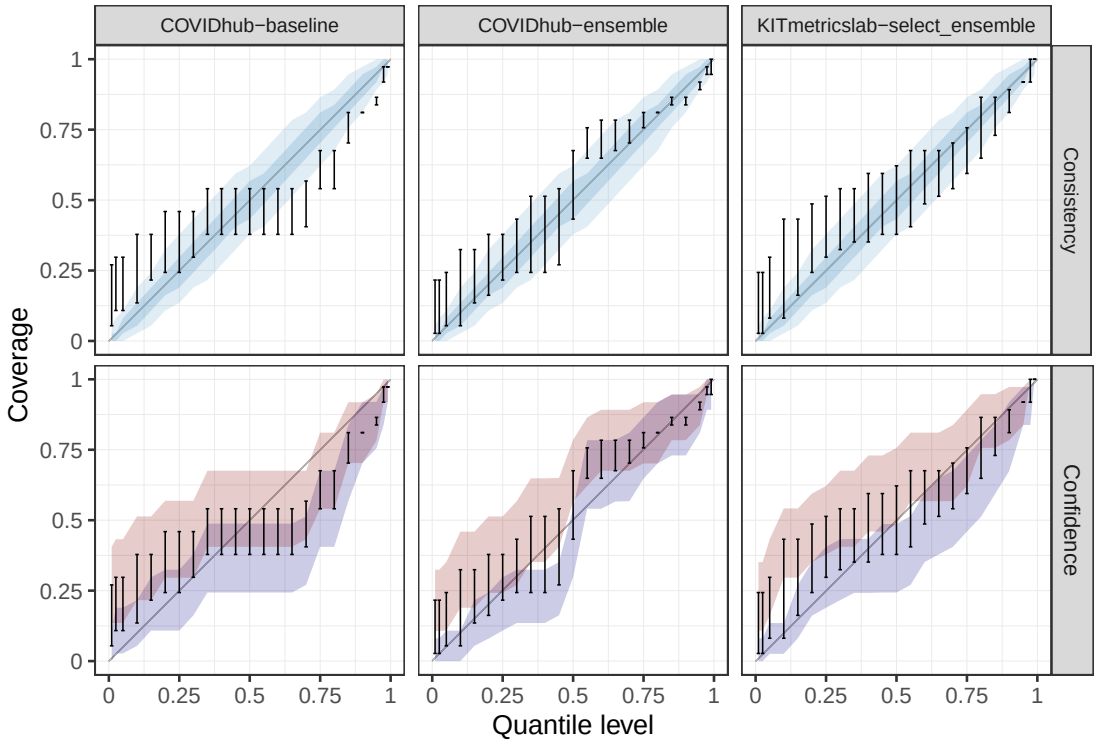


Figure 5.3.: Coverage plots for US COVID-19 Forecast Hub predictions for the state of Vermont. The bars illustrate the interval framed by the lower (c_α^-) and upper (c_α^+) coverage, respectively, which differ substantially. Sampling uncertainty is quantified by 50% and 90% consistency intervals (top row), and 90% confidence intervals for (blue) lower and (red) upper coverage (bottom row).

5.2.3. Conditional calibration

The key diagnostic tool for checking conditional calibration is the quantile reliability diagram, which is an empirical version of the graph introduced in Equation 5.3. To generate such an empirical version, we need to estimate the α-quantile of the outcome as a function of the predicted value. Historically, Bentzien and Friederichs (2014) were the first to propose quantile reliability diagrams, based on bins for the predicted values, and using resampling to quantify uncertainty. The method of Pohle (2020) is based on locally linear kernel regression. We recommend the approach pioneered by Dimitriadis et al. (2021) in the special case of probability forecasts of binary events and by Gneiting and Resin (2023) in the general setting of identifiable statistical functionals.

This approach is based on nonparametric isotonic quantile regression (Wright, 1984) and the pool-adjacent-violators (PAV: Ayer et al., 1955) algorithm for the α-quantile functional (Jordan et al., 2022), to yield statistically Consistent, Optimally binned, Reproducible and PAV-based (CORP) estimates of quantile reliability curves, along with uncertainty quantification via resampling.

For simplicity, suppose without loss of generality that the data of the form in Equation 5.4 satisfy $x_1 \leq x_2 \leq \cdots \leq x_n$. The PAV algorithm generates an increasing sequence

$$\hat{x}_1 \leq \hat{x}_2 \leq \cdots \leq \hat{x}_n \tag{5.6}$$

of (re)calibrated values, which by construction are conditionally α-quantile calibrated with respect to the empirical measure of $(\hat{x}_1, y_1), \ldots, (\hat{x}_n, y_n)$. The algorithm starts with a sequence $\hat{x}_1, \ldots, \hat{x}_n$ of empirical conditional quantiles of the observations given $x \in \{x_1, \ldots, x_n\}$, in the continuous case $\hat{x}_i = y_i$. It proceeds by iteratively pooling adjacent violators of the isotonicity constraint: While there is an index i such that $\hat{x}_i > \hat{x}_{i+1}$, all successive values $\hat{x}_k, \ldots, \hat{x}_l$, where $k = \min\{j \leq i \mid x_j = \cdots = x_i\}$ and $l = \max\{j > i \mid x_{i+1} = \cdots = x_j\}$, are replaced by an empirical quantile of the observations $y_k, \ldots, y_l$. The recalibrated values in Equation 5.6 provide an optimal isotonic estimate under any consistent scoring function, as formally defined in Section 5.3.1 (Gneiting and Resin, 2023; Jordan et al., 2022). The CORP reliability diagram depicts the non-decreasing, piecewise linear function that connects the points $(x_1, \hat{x}_1), \ldots, (x_n, \hat{x}_n)$, and the regularizing constraint of isotonicity leads to optimal binning without the need for tuning parameters. Evidently, the closer the graph to the diagonal, the more desirable.

Consistency bands are constructed from resampled calibration curves in a pointwise fashion, that is, the 90% consistency band at forecast value x spans from the 5th percentile to the 95th percentile of the values attained by the resampled calibration curves at x. Resampled calibration curves are obtained by resampling from the residuals $y_i - x_i$ and adding the resampled residuals to the forecasts while also accounting for unconditional miscalibration to obtain new observations, as described in Gneiting and Resin (2023, Supplement B). Note that statistical inference requires that residuals are independent of calibrated forecasts.

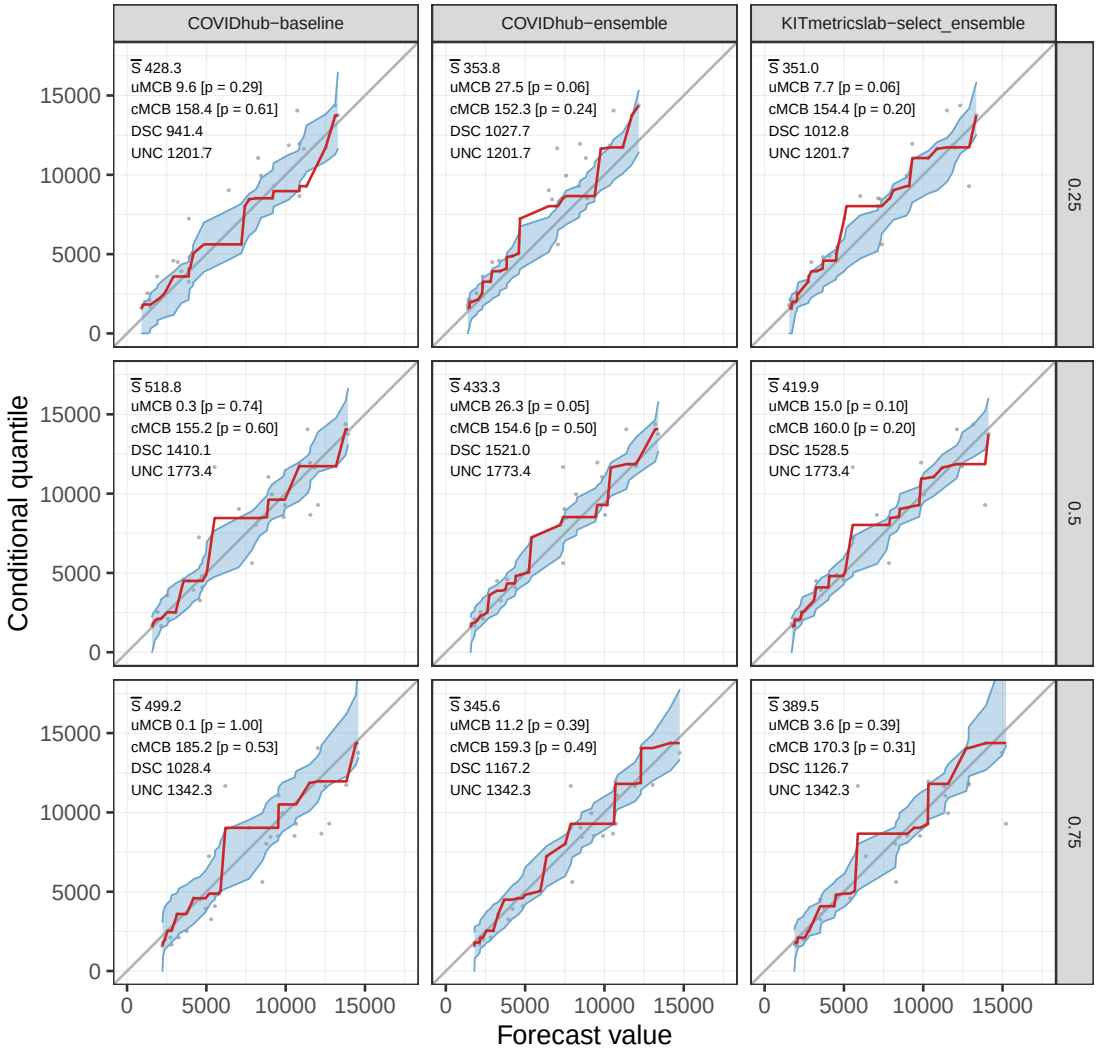


Figure 5.4.: Quantile reliability diagrams ($\alpha = 0.25, 0.5, 0.75$) for US COVID-19 Forecast Hub predictions at the national level, with 90% consistency bands and CORP components of the pinball loss.

For illustration, Figure 5.4 depicts α-quantile reliability diagrams ($\alpha = 0.25, 0.5, 0.75$) for US COVID-19 Forecast Hub predictions at the national level. For all three methods considered, the reliability curves scatter around the diagonal, and the consistency bands suggest compatibility with the assumption of conditional α-quantile calibration. As the inset information suggests, the CORP approach can also be used for score decompositions and tests of conditional calibration, and we refer to Section 5.3.3 for details.

5.3. Comparative evaluation: Consistent scoring functions for quantiles and intervals

Coverage plots and reliability diagrams allow for diagnostic checks of calibration. They serve the purposes of an absolute evaluation, where one seeks to analyze a model's ability to represent and model existing or predict new data. However, many applied settings — including, but by far not limited to, forecast contests — call for relative evaluation, where one seeks to compare and rank competing methods in terms of their performance. Consistent scoring functions are tailored to this task, and we turn to them now.

5.3.1. Consistent scoring functions for quantiles

A loss or scoring function is a function $\mathrm{S} : \mathrm{D} \times \mathrm{D} \to [0, \infty)$ in two arguments, where $\mathrm{S}(x, y)$ is interpreted as the loss or penalty when $x \in \mathrm{D}$ is the forecast and $y \in \mathrm{D}$ the outcome. We assume throughout that the loss under a perfect forecast vanishes, that is, $\mathrm{S}(x, x) = 0$ for $x \in \mathrm{D}$, and we suppose that the outcome domain D of interest is either the real line, $\mathrm{D} = \mathbb{R}$, or the positive halfaxis, $\mathrm{D} = (0, \infty)$. Following Gneiting (2011a), a scoring function S is consistent for the α-quantile if, given any cumulative distribution function F on the domain D and any $t \in Q_\alpha(F) \subseteq \mathrm{D}$,

$$\mathbb{E}_F \mathrm{S}(t, Y) \leq \mathbb{E}_F \mathrm{S}(x, Y) \quad \text{for all} \quad x \in \mathrm{D}.$$

It is strictly consistent if the inequality is strict unless $x \in Q_\alpha(F)$. Hence, under a consistent loss or scoring function, the α-quantile functional serves as Bayes rule or optimal point forecast. In this light, consistent scoring functions are incentive-compatible and elicit the α-quantile. In other words, the use of consistent scoring functions rewards (respectively, penalizes) modelers and forecasters, and encourages honest and careful assessments.

Subject to minor regularity conditions, a scoring function is consistent for the α-quantile if, and only if, it is of the generalized piecewise linear (GPL) form, that is,

$$\mathrm{S}(x, y) = (\mathbb{1}\{y \leq x\} - \alpha)\,(g(x) - g(y)) \tag{5.7}$$

for some nondecreasing function g on D. If g is strictly increasing, then S is strictly consistent subject to integrability constraints. An equivalent characterization is due to Thomson (1979), though the GPL form in Equation 5.7 emerged considerably later only (Saerens, 2000; Gneiting, 2011a; Grant and Gneiting, 2013).[2] The most prominent example is the popular asymmetric piecewise linear scoring function or pinball loss,

$$\mathrm{S}_\alpha(x, y) = (\mathbb{1}\{y \le x\} - \alpha)\,(x - y) = \begin{cases} (1-\alpha)\,(x-y), & y \le x, \\ \alpha\,(y-x), & y \ge x, \end{cases} \tag{5.8}$$

which is of the prediction error form, that is, it depends on the residual $x - y$ only, and serves as ubiquitous loss function in estimating quantile regression models (Koenker and Bassett, 1978; Koenker, 2005, 2017) and in scoring forecast contests (Chen et al., 2022). Gneiting and Resin (2023) refer to S_α as the canonical loss or scoring function for the α-quantile functional. When $\alpha = 0.5$ we recover the absolute error, up to a constant factor.[3] Interestingly, if a quantile regression model includes an intercept, in-sample score optimization with respect to the canonical loss is essentially equivalent to enforcing unconditional calibration (Gneiting and Resin, 2023, Theorem 2.26), with the partitioning inequalities of linear quantile regression (Koenker and Bassett, 1978, Theorem 3.4) being an immediate consequence.

In applied work, scoring functions with equivariance properties are often preferred. Nolde and Ziegel (2017, Supplement C) characterize the respective members of the GPL class. In particular, if $\mathrm{D} = (0, \infty)$ there is a unique (up to a constant multiple) scale invariant member of the GPL class, namely,

$$\mathrm{S}(x, y) = (\mathbb{1}\{y \le x\} - \alpha) \log \frac{x}{y}.$$

[2] For quantile forecasts and their evaluation, we have to interpret the observed outcome as the true target variable, even when it is subject to measurement error, as the GPL form of Equation 5.7 does not allow otherwise (Hoga and Dimitriadis, 2022, Supplement A). This is in contrast to mean forecasts, where robust (Bregman) loss functions yield correct forecast rankings even when the observed outcome is subject to (conditionally unbiased) measurement error (Patton, 2011).

[3] Evidently, S_α fails to be differentiable on the diagonal. In this light, smooth approximations have been proposed, see, for example, Fasiolo et al. (2021). However, approximations may not yield consistent scoring functions.

As noted, a widely used form of a nominal $(100 \times \beta)\%$ interval is bounded by predictive α-quantiles, where $\alpha = \frac{1}{2}(1 - \beta)$ and $\alpha = \frac{1}{2}(1 + \beta)$, respectively.[4] The key example of a consistent scoring function for this type of interval forecast arises when one adds the respective pinball losses S_α from Equation 5.8. This yields Winkler's interval score (Winkler, 1972; Gneiting and Raftery, 2007), which has intuitively appealing interpretations in terms of the forecast goals, namely, the maximization of the sharpness of the predictive interval subject to calibration (Gneiting et al., 2007; Gneiting and Katzfuss, 2014).

5.3.2. Mixture representations and Murphy diagrams

A useful mixture representation for the members of the GPL class was introduced by Ehm et al. (2016, Theorem 1a). Subject to minor regularity conditions, any consistent scoring function for the α-quantile admits a representation of the form

$$\mathrm{S}(x, y) = \int_{-\infty}^{+\infty} \mathrm{S}_{\alpha,\theta}(x, y)\ \mathrm{d}H(\theta) \tag{5.9}$$

where H is a nonnegative measure and

$$\mathrm{S}_{\alpha,\theta}(x, y) = (\mathbb{1}\{y \le x\} - \alpha)\ (\mathbb{1}\{x > \theta\} - \mathbb{1}\{y > \theta\}) = \begin{cases} 1 - \alpha, & y \le \theta < x, \\ \alpha, & x \le \theta < y, \\ 0, & \text{otherwise}, \end{cases} \tag{5.10}$$

is the elementary quantile scoring function at level $\alpha \in (0, 1)$ and threshold $\theta \in \mathbb{R}$.[5] The mixing measure H is unique and satisfies $\mathrm{d}H(\theta) = \mathrm{d}g(\theta)$, where g is the nondecreasing function in Equation 5.7. For example, if H is the Lebesgue measure, Equation 5.9 recovers the canonical pinball loss of Equation 5.8. As an immediate consequence of the

[4] For recent progress in the evaluation of interval forecasts see Brehmer and Gneiting (2021), Fissler et al. (2021), Taylor (2021) and references therein. Also, scoring rules for predictive quantiles at several levels can be combined (Jose and Winkler, 2009). Bracher et al. (2021a) developed the weighted interval score that considers the 23 levels at the US COVID-19 Forecast Hub, and the GEFCom14 competition used a sum of pinball losses at 99 levels (Hong et al., 2016).

[5] Equation 5.9 can be interpreted as a Choquet representation in the sense of functional analysis, with the $\mathrm{S}_{\alpha,\theta}$ in Equation 5.10 being the elementary or extreme members of the convex cone of the consistent scoring functions (Ehm et al., 2016).

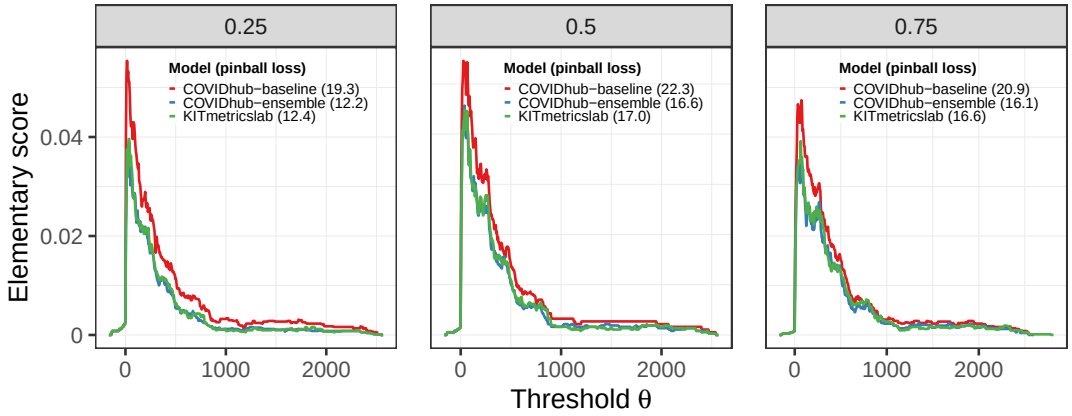


Figure 5.5.: Murphy diagrams for US COVID-19 Forecast Hub α-quantile forecasts ($\alpha = 0.25, 0.5, 0.75$) at the state level.

mixture representation, a forecast method is preferable over another in terms of just any consistent scoring function if, and only if, it is preferable in terms of the elementary scoring functions.

Let us turn attention to the empirical setting and the mean empirical score,

$$\bar{\mathrm{S}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(x_i, y_i), \tag{5.11}$$

for data of the form in Equation 5.4. In a Murphy diagram, we plot the graphs of the mean elementary quantile score $\mathrm{S}_{\alpha,\theta}$, namely, the Murphy curve $\theta \mapsto \bar{\mathrm{S}}_{\alpha,\theta}$ for competing forecasting methods. The area under its Murphy curve equals a method's mean pinball loss. If a method has a Murphy curve entirely below the Murphy curve of another method, then it is preferable in terms of any consistent scoring function. If the Murphy curves cross, there is a lack of dominance and there may not be a clear-cut preference (Ehm et al., 2016). For illustration, Figure 5.5 shows Murphy diagrams for three models from the US COVID-19 Forecast Hub. At all levels considered, the COVIDhub-ensemble and KITmetrics-select_ensemble models outperform the baseline.

While Murphy diagrams serve as diagnostic tools, applications frequently call for inference about predictive performance. In such settings, the popular Diebold and

Mariano (1995) test of equal predictive performance applies, which compares empirical scores of the form in Equation 5.11 in time series settings. A detailed analysis for quantile forecasts under the pinball loss was provided by Giacomini and Komunjer (2005). For a recent perspective on testing, with particular emphasis on sequential tests, see Choe and Ramdas (2021) and references therein.

5.3.3. CORP decomposition

It is often desirable to decompose an empirical score of the form in Equation 5.11 into interpretable components. We follow Gneiting and Resin (2023) and describe such a decomposition that is based on the aforementioned CORP approach.

Starting from data of the form in Equation 5.4, suppose again that $x_1 \leq \cdots \leq x_n$ and let $\hat{x}_1 \leq \ldots \leq \hat{x}_n$ denote the conditionally (and unconditionally) α-quantile calibrated values from Equation 5.6 as generated by the PAV algorithm. Furthermore, let $\hat{x}_0$ be the α-quantile of the outcomes $y_1, \ldots, y_n$. If S is a consistent scoring function, let

$$\bar{\mathrm{S}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(x_i, y_i), \quad \bar{\mathrm{S}}_{\mathrm{rc}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(\hat{x}_i, y_i), \quad \text{and} \quad \bar{\mathrm{S}}_{\mathrm{mg}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(\hat{x}_0, y_i)$$

denote the mean score of the posited predictive quantile, of its (re)calibrated version, and of the unconditional or marginal quantile. We refer to

$$\mathsf{MCB} = \bar{\mathrm{S}} - \bar{\mathrm{S}}_{\mathrm{rc}}, \quad \mathsf{DSC} = \bar{\mathrm{S}}_{\mathrm{mg}} - \bar{\mathrm{S}}_{\mathrm{rc}}, \quad \text{and} \quad \mathsf{UNC} = \bar{\mathrm{S}}_{\mathrm{mg}}$$

as the miscalibration (MCB), discrimination (DSC) and uncertainty (UNC) components of the mean score $\bar{\mathrm{S}}$, all of which are nonnegative by construction. The mean score $\bar{\mathrm{S}}$ then decomposes into a signed sum of readily interpretable components,

$$\bar{\mathrm{S}} = \mathsf{MCB} - \mathsf{DSC} + \mathsf{UNC}, \tag{5.12}$$

where, notably, $\mathsf{MCB} \geq 0$ with equality if $\hat{x}_i = x_i$ for $i = 1, \ldots, n$, and $\mathsf{DSC} \geq 0$ with equality if $\hat{x}_i = \hat{x}_0$ for $i = 1, \ldots, n$. If S is strictly consistent, then $\mathsf{MCB} = 0$ only if $\hat{x}_i = x_i$ for $i = 1, \ldots, n$ and $\mathsf{DSC} = 0$ only if $\hat{x}_i = \hat{x}_0$ for $i = 1, \ldots, n$. The MCB component equals the difference in mean score of the predictive quantiles at hand and their calibrated

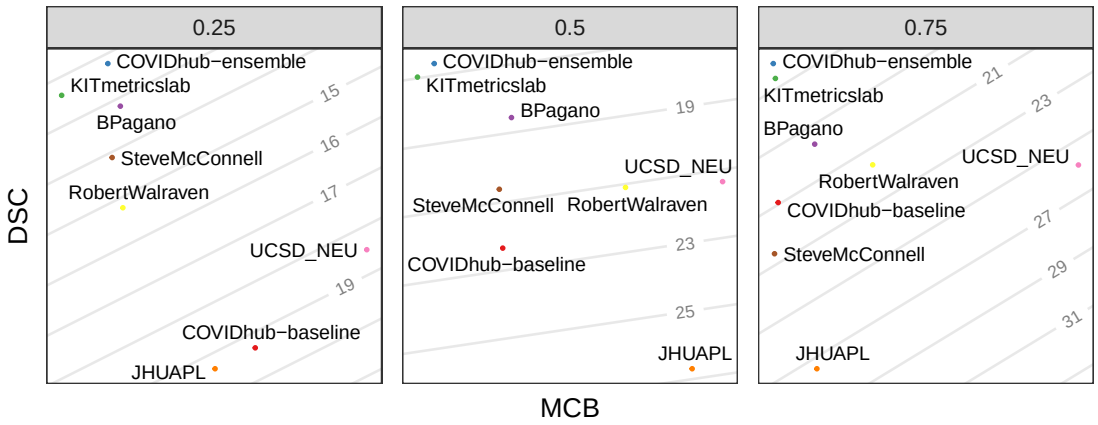


Figure 5.6.: CORP components of pinball loss for US COVID-19 Forecast Hub α-quantile forecasts ($\alpha = 0.25, 0.5, 0.95$) at the state level. Pinball loss is constant along the light gray lines.

version and quantifies miscalibration. The DSC component equals the difference in mean score to a calibrated but constant forecast and is a measure of discrimination ability. The UNC component depends on the outcomes only.

If more than a handful of forecasting methods are to be compared, the use of coverage plots, reliability diagrams, and Murphy diagrams may become cumbersome. In such settings, it can be instructive to consider the CORP decomposition of the pinball loss and plot the DSC versus the MCB component. For an illustration on US COVID-19 Forecast Hub models see Figure 5.6, where the diagonal lines indicate equal mean pinball loss. Methods that appear at top left show the smallest mean score. The more left a model the better it is calibrated, and the higher a model the better it discriminates. For example, we see that at all levels considered, the COVIDhub-ensemble discrimates best. The KITmetrics-select_ensemble is very well calibrated, but lags behind the COVIDhub-ensemble in discrimination ability.

For a refinement of the CORP decomposition in Equation 5.12 in the case of the canonical pinball loss,[6] let d_α be such that the empirical measure in $(x_1 + d_\alpha, y_1), \ldots, (x_n +$

[6]For the extended decomposition in Equation 5.13 we restrict attention to the pinball loss, as the crucial property that $\mathsf{MCB}_\mathrm{u} \geq 0$ might be violated under non-canonical GPL functions.

$d_\alpha, y_n)$ is unconditionally α-quantile calibrated, and define

$$\bar{\mathrm{S}}_{\mathrm{urc}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(x_i + d_\alpha, y_i)$$

as the mean score of the unconditionally recalibrated α-quantile forecast. The unconditional ($\mathsf{MCB_u}$) and conditional ($\mathsf{MCB_c}$) miscalibration components are given by

$$\mathsf{MCB_u} = \bar{\mathrm{S}} - \bar{\mathrm{S}}_{\mathrm{urc}} \quad \text{and} \quad \mathsf{MCB_c} = \bar{\mathrm{S}}_{\mathrm{urc}} - \bar{\mathrm{S}}_{\mathrm{rc}},$$

and it holds that $\mathsf{MCB} = \mathsf{MCB_u} + \mathsf{MCB_c}$, where $\mathsf{MCB_u} \geq 0$ and $\mathsf{MCB_c} \geq 0$. We then note the extended CORP decomposition,

$$\bar{\mathrm{S}} = \mathsf{MCB_u} + \mathsf{MCB_c} - \mathsf{DSC} + \mathsf{UNC}. \tag{5.13}$$

This is the decomposition of the pinball loss that we show in the reliability diagrams and table in the paper. The p-value for unconditional α-quantile calibration stems from the binomial test described in Section 5.2.2, and the p-value for conditional α-calibration uses resampling and the $\mathsf{MCB_c}$ component as test statistics, as described in Section 5.2.3 and Gneiting and Resin (2023, Supplement B).

5.3.4. Skill scores and coefficient of determination

In out-of-sample forecast evaluation, the quantity

$$\mathrm{S}_{\mathrm{skill}} = 1 - \frac{\bar{\mathrm{S}}}{\bar{\mathrm{S}}_{\mathrm{mg}}} = \frac{\bar{\mathrm{S}}_{\mathrm{mg}} - \bar{\mathrm{S}}}{\bar{\mathrm{S}}_{\mathrm{mg}}} \tag{5.14}$$

is known as skill score (Murphy and Epstein, 1989; Gneiting and Raftery, 2007), with positive values indicating better performance than the uninformative baseline, and negative values worse performance.

In contrast, for in-sample model diagnostics the quantity in Equation 5.14 typically is nonnegative (Gneiting and Resin, 2023, Theorem 3.6), and when specialized to the pinball loss it reduces to the coefficient of determination or R^1 measure introduced by Koenker and Machado (1999) and studied by Noh et al. (2013) as a goodness of fit

measure in quantile regression. The CORP decomposition of Equation 5.12 allows for an interesting interpretation, in that $S_{\text{skill}} = R^1 = (\mathsf{DSC} - \mathsf{MCB})/\mathsf{UNC}$. The quantity is dimensionless and, subject to modest regularity conditions, takes values in the unit interval, with a value of 1 indicating a perfect fit, and a value of 0 suggesting lack of fit.

5.4. Empirical examples

We illustrate the use of coverage plots, reliability diagrams, scores and score components on Engel's food expenditure data and the Global Energy Forecasting Competition 2014.

5.4.1. Engel's food expenditure data: In-sample regression diagnostics vs. out-of-sample forecast evaluation

We consider quantile regression fits for the classical food expenditure data from Engel (1857) for 19th century European working class households, as also discussed by Koenker (2005, pp. 78, 297–307). Engel's conclusion that the share of income that is used for food expenditure decreases with income is known as Engel's law and stands in today's work on poverty and especially poverty reduction as one of the most enduring relationships in economics (Blundell et al., 2007). Modeling conditional quantiles for a range of levels ($\alpha = 0.1, 0.25, 0.5, 0.75, 0.9$), instead of the conditional mean, allows a comprehensive assessment of Engel's law. We compare standard (linear) quantile regression, linear quantile regression on log-transformed values (Koenker, 2005, p. 78) and nonparametric isotonic quantile regression (Wright, 1984) both in-sample and out-of-sample, using leave-one-out cross-validation, based on Engel's data of size $n = 235$.

We first fit a standard quantile regression model of food expenditure on income. The parametric form imposes a linear relationship, which in view of Engel's law is too restrictive. Following Koenker (2005, p. 78), we also use a linear model of the log-transformed quantities, where slope values smaller than one support Engel's law, for $\log(y) = \beta_1 + \beta_2 \log(x)$ is equivalent to $y = \exp(\beta_1)x^{\beta_2}$, so that $\beta_2 < 1$ implies a concave relationship. Indeed, we find estimated slope coefficients between 0.80 and 0.92. Finally, we use isotonic quantile regression as a fully flexible nonparametric method. Evidently, the isotonicity assumption is satisfied.

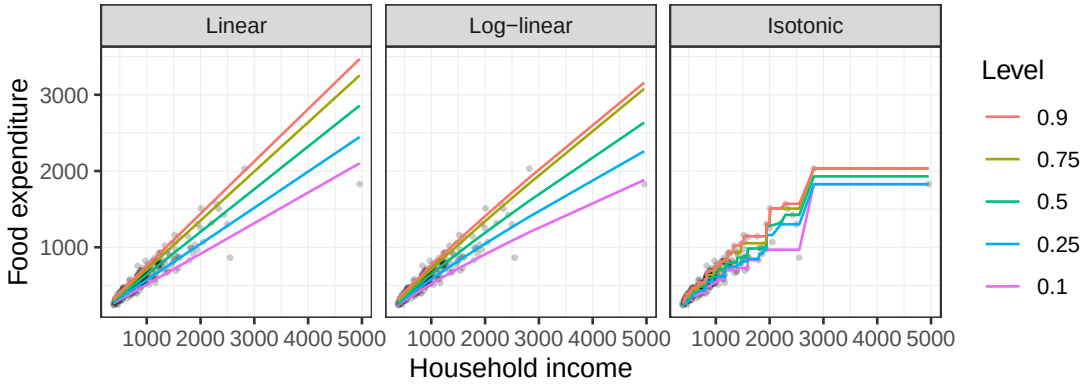


Figure 5.7.: Linear, log-linear and isotonic quantile regression fits for Engel (1857) food expenditure data against household income.

Figure 5.7 shows the in-sample model fit for the three methods and five quantile levels. The log-linear model fits show a slightly concave shape, as can be expected by Engel's law, which is confirmed by the nonparametric isotonic estimates.

In Figure 5.8, Figure 5.9 and Table 5.1, we contrast in-sample model diagnostics and out-of-sample (leave-one-out cross-validation) forecast evaluation for the three methods.[7] Perfect in-sample coverage is guaranteed by the partitioning inequalities of quantile regression. Similarly, isotonic regression fits show perfect in-sample unconditional and conditional calibration by construction, with reliability diagrams that are constrained to the diagonal. While the linear and log-linear models retain good coverage out-of-sample, unconditional and conditional calibration deteriorate notably for the isotonic model.

[7] To generate consistency bands in the reliability diagrams, we resample residuals of log-transformed values, which seems natural here and in other applications with strictly positive data, where variability increases as observed values increase.

Table 5.1.: In-sample and out-of-sample CORP components of the pinball loss $\bar{\mathrm{S}}$ for α-quantile regression fits to Engel (1857) food expenditure data

Level (UNC)	Component	In-sample			Out-of-sample		
		Linear	Log-Linear	Isotonic	Linear	Log-Linear	Isotonic
$\alpha = 0.1$ (32.6)	$\bar{\mathrm{S}}$	16.5	15.1	12.0	17.5	15.2	18.3
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.4
	$\mathsf{MCB_c}$	4.5	3.1	0.0	5.5	3.2	4.7
	DSC	20.6	20.6	20.6	20.6	20.6	19.4
$\alpha = 0.25$ (67.6)	$\bar{\mathrm{S}}$	30.1	29.2	23.0	30.6	29.8	30.8
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.1
	$\mathsf{MCB_c}$	7.1	6.2	0.0	7.3	6.6	4.4
	DSC	44.6	44.6	44.6	44.3	44.4	41.3
$\alpha = 0.5$ (98.5)	$\bar{\mathrm{S}}$	37.4	36.6	28.5	38.0	37.3	37.8
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathsf{MCB_c}$	8.9	8.1	0.0	8.9	8.1	5.5
	DSC	70.0	70.0	70.0	69.4	69.3	66.2
$\alpha = 0.75$ (91.6)	$\bar{\mathrm{S}}$	27.9	27.6	21.0	28.6	28.4	30.0
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathsf{MCB_c}$	6.9	6.6	0.0	6.9	6.6	5.2
	DSC	70.6	70.6	70.6	69.9	69.8	66.8
$\alpha = 0.9$ (61.3)	$\bar{\mathrm{S}}$	14.4	14.4	10.2	15.0	14.9	17.8
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.6
	$\mathsf{MCB_c}$	4.2	4.2	0.0	4.4	4.3	5.4
	DSC	51.1	51.1	51.1	50.7	50.7	49.5

Table 5.1 shows the mean pinball loss along with the CORP decomposition. In-sample, all three methods show perfect unconditional calibration with a vanishing $\mathsf{MCB_u}$ component, and they also share the DSC component, for they are isotonic transformations of each other. The $\mathsf{MCB_c}$ component vanishes for the isotonic model and is slightly better for the log-linear than for the linear model. The nonparametric isotonic regression technique is prone to overfitting in small samples, which results in the best scores in-sample, but worse scores out-of-sample when compared to the parametric linear and log-linear models. Interestingly, isotonic regression also has the worst out-of-sample DSC components. Both in-sample and out-of-sample, the log-linear model has slightly better scores than the linear model, providing additional support for Engel's law.

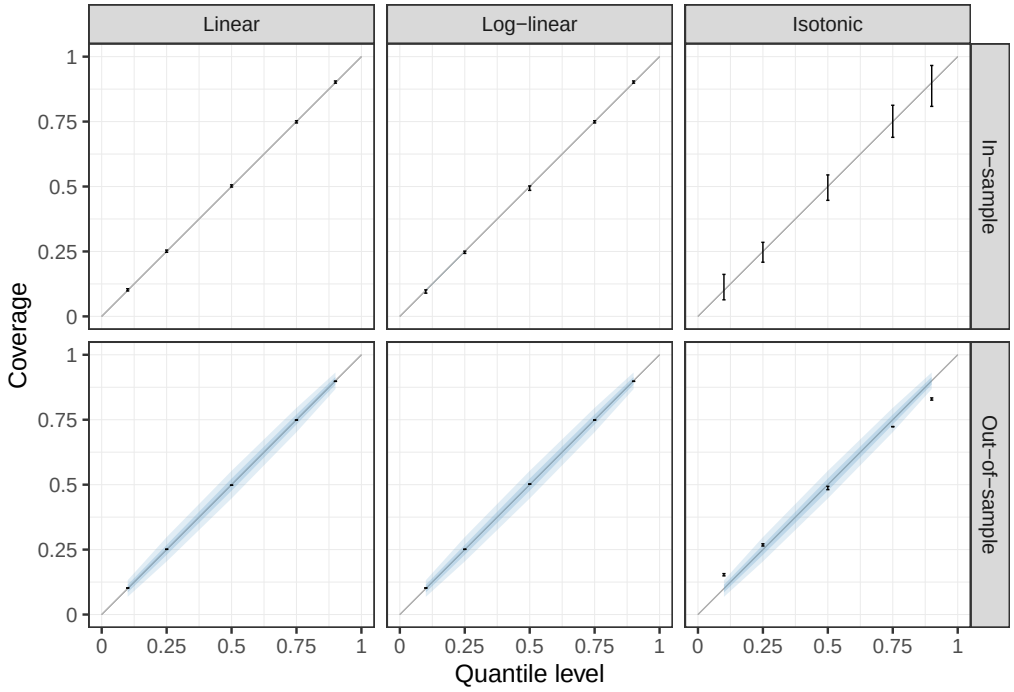


Figure 5.8.: In-sample (top row) and out-of-sample (bottom row) coverage plots, depicting the intervals $[c_\alpha^-, c_\alpha^+]$ along with 50% and 90% consistency bands, for quantile regression fits to Engel (1857) food expenditure data.

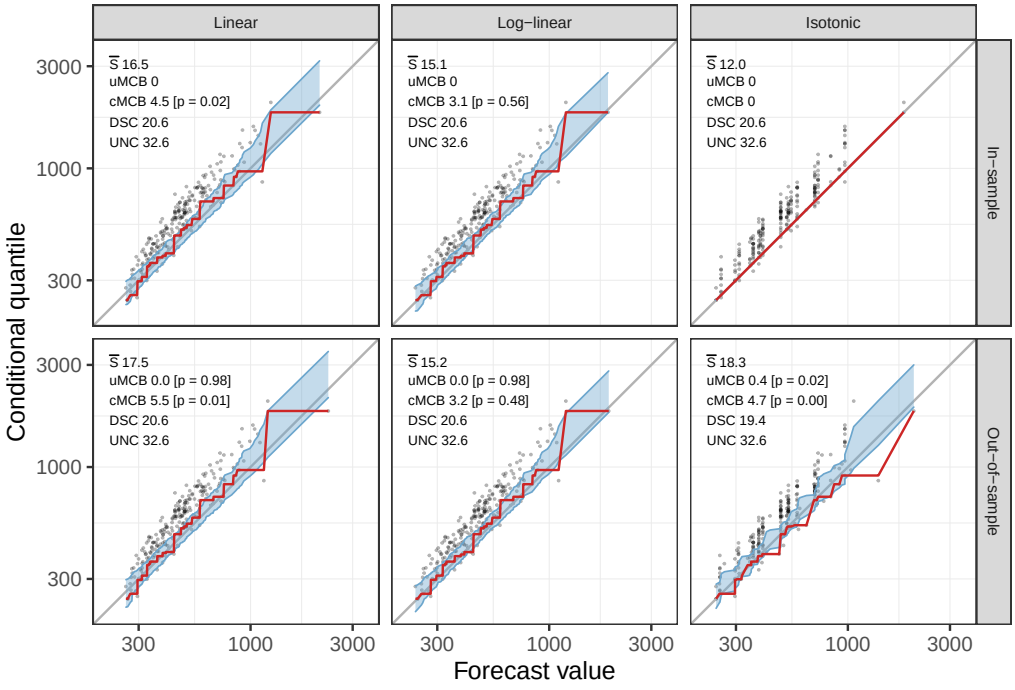


Figure 5.9.: In-sample (top row) and out-of-sample (bottom row) α-quantile reliability diagrams ($\alpha = 0.1$), along with 90% consistency bands, for quantile regression fits to Engel (1857) data.

5.4.2. Global Energy Forecasting Competition 2014

For a successful transition to an energy system characterized by a high share of renewable energy sources, such as wind and solar power, accurate and reliable forecasts of energy supply, demand, and prices are paramount. Furthermore, these forecasts should be probabilistic in nature to facilitate optimal decision making for planning and operations in the energy system (Hong et al., 2020). The Global Energy Forecasting Competition 2014 (GEFCom2014; Hong et al., 2016) sought to foster research on such probabilistic forecasts.

To illustrate the use of quantile forecast diagnostics on energy system data, we consider the GEFCom2014 wind power forecasting track, where the aim was to predict wind power generation for 10 zones, corresponding to wind farms in Australia at undisclosed locations. The wind power values were provided on a normalized scale with proportions of the nominal capacity of the wind farm between 0 and 1. Available predictors included wind forecasts for the exact location of the wind farms from the European Centre for Medium-range Weather Forecasts at heights of 10 and 100 m above ground in the form of the zonal and meridional wind components. The wind forecasts were available both for training and as inputs over the out-of-sample evaluation period of the various tasks. Hourly forecasts were to be submitted on a rolling basis with forecast lead times up to 24 hours ahead, starting each day at midnight. The 15 tasks during the competition period covered one month of data each. In the following, we consider all tasks but restrict our attention to Zone 1. The predictions were to be provided in the form of 99 predictive quantiles at levels 0.01, 0.02, ..., 0.98, and 0.99.

In Figure 5.10, we show forecasts based on isotonic distributional regression (IDR; Henzi et al., 2021), the nearest neighbor quantile filter (NNQF; González Ordiano et al., 2020), and quantile regression forests (QRF; Meinshausen, 2006). For IDR we use the predicted wind speed at 100 m height as single explanatory variable. To refine this approach, we stratify and assign each instance to a group, determined by eight equally spaced bins of the forecasted wind direction. The IDR cond method trains and applies IDR on the respective groups, using an implementation in R (Henzi, 2021; R Core Team, 2021). The NNQF + MLP method uses the NNQF in concert with a Multi-Layer Perceptron (MLP) regressor for each quantile level, implemented in Python (Pedregosa et al., 2011; Python

Software Foundation, 2022). As features, we use scaled (to variance one) forecasts of wind components and wind speed at lags of zero up to 11 hours ago. The preprocessing by NNQF uses the Euclidian distance and 50 nearest neighbors, based on non-lagged features only. The MLP regressors use lagged features and have a single hidden layer with 100 units. To obtain valid conditional quantiles of the normalized response variable, we replace negative values with 0 and predictions larger than 1 with 1. For QRF, we use default options in the implementation by Wright and Ziegler (2017).

For the top row, the predictive quantiles at levels 0.05, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, and 0.95 are extracted to show the predictive median and centered prediction intervals with nominal level at 20%, 50%, 80%, and 90%, respectively. The coverage plots and reliability diagrams ($\alpha = 0.75$) demonstrate that although the QRF method has the highest discrimination ability and the lowest pinball loss, all three methods only deviate modestly from unconditional calibration and conditional α-quantile calibration.[8] The Murphy diagram ($\alpha = 0.75$) supports these findings and suggests that essentially all forecast users will be served best by using the QRF method. However, the differences in performance between the three methods pale when compared to the original contributions to GEFCom2014. According to the GEFCom2014 leaderboard, all three methods would have ranked similarly between the sixth and the eighth best entry in the competition (Hong et al., 2016, Supplement).

[8] Caution is warranted when interpreting the consistency bands in the reliability diagrams. These assume exchangeability of the forecast situations (Gneiting and Resin, 2023, Supplement B), a condition that is violated here, as the contest format incurs temporal dependence. Thus, the consistency bands seem unrealistically narrow. For the coverage plots, we do not show consistency or confidence bands.

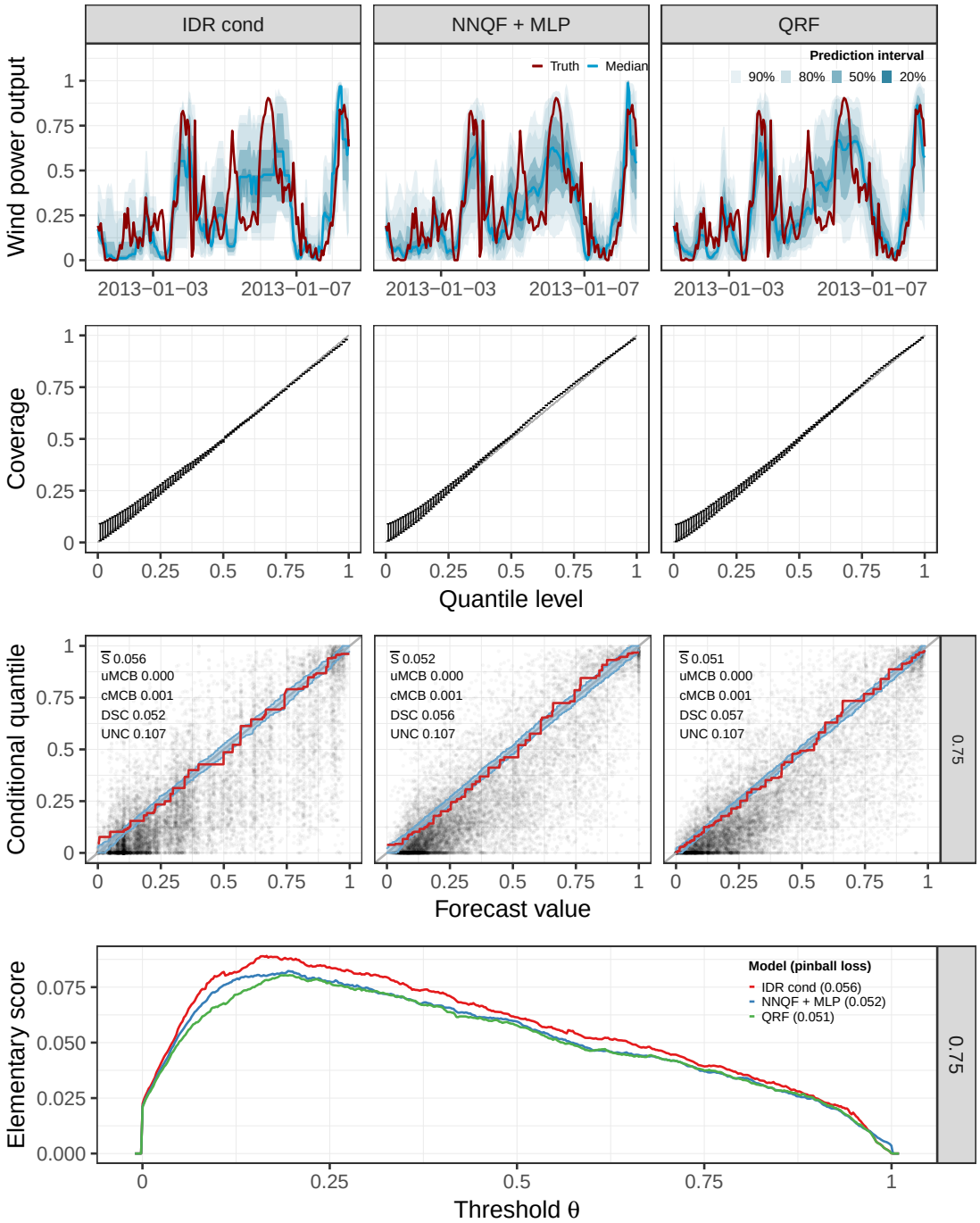


Figure 5.10.: Quantile forecasts, coverage plots, α-quantile reliability diagrams ($\alpha = 0.75$) and Murphy diagram ($\alpha = 0.75$) for IDR, NNQF + MLP and QRF based entries to the wind power track (Zone 1) in the GEFCom2014 contest.

5.5. Discussion

We reviewed tools for in-sample diagnostics and out-of-sample forecast evaluation of quantile models. Calibration concerns the statistical compatibility between the posited predictive quantiles and outcomes. Unconditional calibration corresponds to classical coverage criteria, and the stronger notion of conditional calibration can be diagnosed in quantile reliability diagrams. Adequate handling of discrete data is critical, particularly in checking coverage. Consistent scoring functions allow for comparative assessment and ranking, with the canonical piecewise linear or pinball loss being the most prominent example. The recently developed CORP approach allows for regularized, isotonic estimates of reliability curves, and admits decompositions of the pinball loss and other consistent scoring functions into interpretable components.

The methods reviewed are entirely generic and can be applied in any context where predictive statements in the form of quantiles are to be assessed, be it in time series format, spatial or spatio-temporal settings, in clinical trials, or any other type of setting. Of course, stronger methods might become available if structure can be exploited; for example, in sequential time series forecasts for one step ahead, in addition to the predictive quantiles being calibrated, exceedances ought to be independent (Christoffersen, 1998).

The notions and tools presented here adhere to the prequential principle (Dawid, 1984), in that they depend on data of the form in Equation 5.4 only, but not on the way that the predictive quantiles were generated. Stronger notions of calibration may also be useful, for example, conditionally on features or covariates (Chung et al., 2021; Fissler and Pesenti, 2023).

In this paper, we focused attention on quantiles at moderate levels. While quantiles at extreme levels play crucial roles in practice, the evaluation of predictive performance for extreme quantiles, and extreme events in general, is subject to further challenges (Lerch et al., 2017) and ongoing activity (Brehmer and Strokorb, 2019; Gandy et al., 2022; Taggart, 2022). The adequate handling of censored data in survival analysis and related types of applications is another recently studied area (Li and Peng, 2017; De Backer et al., 2019; Conde-Amboage et al., 2021) that calls for follow-up work.

Finally, while the methods presented serve diagnostic purposes well, tools for statistical inference about predictive performance, such as the generation of consistency and confi-

dence bands, and the development of hypothesis tests, require further development. This is especially the case in structured data settings including, but not limited to, the much studied time series case (Christoffersen, 1998; Giacomini and Komunjer, 2005; Nolde and Ziegel, 2017). In the CORP framework, methods for the generation of consistency and confidence bands based on asymptotic studies (Wright, 1984; Mösching and Dümbgen, 2020) might complement currently available, resampling-based methods (Gneiting and Resin, 2023, Supplement B).

6. Shift-dispersion decompositions of Wasserstein and Cramér distances

6.1. Introduction

The task of comparing pairs of probability distributions arises in numerous contexts and has given rise to a wealth of *divergence functions* or *statistical distances* (Deza and Deza, 2013). A particularly well-studied divergence is the Wasserstein distance with numerous theoretical developments in machine learning (e.g., Frogner et al. 2015; Arjovsky et al. 2017), dependence modelling (Wiesel, 2022), distributional regression (Chen et al., 2023b), and model diagnostics (Munk and Czado, 1998). Applications exist in diverse fields including image processing (Ni et al., 2009), biostatistics (Schefzik et al., 2021), and economics (Gini, 1914; Rachev et al., 2011); see Panaretos and Zemel (2019) for a recent review. The Cramér distance is an alternative which shares many attractive properties of the Wasserstein distance, including its symmetry property and the rewarding of closeness (Rizzo and Székely, 2016). It has become a popular metric in forecast evaluation, with applications in the atmospheric sciences (Thorarinsdottir et al., 2013; Richardson et al., 2020), hydrology (Barna et al., 2023) and electricity markets (Janke and Steinke, 2019). The Cramér distance has also found use in machine learning as an alternative to the Wasserstein distance in generative adversarial networks (Bellemare et al., 2017).

In this paper, we are concerned with decompositions of statistical distances of Wasserstein and Cramér type. We focus on the real-valued case, where F and G are probability distributions on the real line $\mathbb{R}$, and we identify both distributions with their respective cumulative distribution functions (CDFs). In a nutshell, a divergence is a function D such that $\mathrm{D}(F, G)$ is non-negative for all F, G, and equals zero if and only if $F = G$. The purpose of divergence functions is to reduce the difference between two probability

distributions, i.e., two infinite-dimensional objects, to a single number. Of course, this reduction implies a severe loss of information: A divergence function only quantifies the *magnitude of dissimilarity* between two distributions F and G, but it hides the *specific nature of the differences*, e.g., whether the main difference is in location or dispersion. To shed light on these aspects, we propose novel decompositions of divergence functions D into four non-negative and interpretable components

$$\mathrm{D}(F,G) = \mathrm{Shift}_+^{\mathrm{D}} + \mathrm{Shift}_-^{\mathrm{D}} + \mathrm{Disp}_+^{\mathrm{D}} + \mathrm{Disp}_-^{\mathrm{D}}\,. \tag{6.1}$$

Here, the *shift components* $\mathrm{Shift}_\pm^{\mathrm{D}}$ with $\pm \in \{+,-\}$ quantify differences in *location*, while the *dispersion components* $\mathrm{Disp}_\pm^{\mathrm{D}}$ measure differences in *variability* between F and G. The signed components (with subscript '+' and '−') attribute parts of the distance to upwards and downwards shifts, and more or less dispersion of F relative to G. Of course, the components in (6.1) are functions of the pair of distributions (F,G). For the sake of brevity, however, we will sometimes omit this dependence and use the shorthands $\mathrm{Shift}_\pm^{\mathrm{D}} = \mathrm{Shift}_\pm^{\mathrm{D}}(F,G)$ and $\mathrm{Disp}_\pm^{\mathrm{D}} = \mathrm{Disp}_\pm^{\mathrm{D}}(F,G)$. These refer to the components of $\mathrm{D}(F,G)$ between for two generic distributions F and G or a pair of distributions which becomes clear from the context.

Our decompositions in (6.1) apply to arbitrary (possibly discontinuous) distributions and to divergence measures that allow for certain representations through quantile functions. We address the aforementioned Wasserstein distance, more specifically, the p-th power of the *p-Wasserstein distance*,

$$\mathrm{WD}_p(F,G) = \int_0^1 \left|F^{-1}(\tau) - G^{-1}(\tau)\right|^p \mathrm{d}\tau \tag{6.2}$$

for $p \in \mathbb{N}$. Here, F^{-1} and G^{-1} denote the generalized quantile functions, given by $F^{-1}(\tau) = \inf\{x \in \mathbb{R} \mid \tau \leq F(x)\}$, $\tau \in [0,1]$ and accordingly for G^{-1}, i.e., the left-inverses of the CDFs. The *Cramér distance* arises as the square of the special case $p = 2$ within the class of l_p distances,

$$l_p(F,G) = \left(\int_{-\infty}^{\infty} |F(x) - G(x)|^p \,\mathrm{d}x\right)^{1/p}. \tag{6.3}$$

While its classic definition (as a special case of (6.3)) is in terms of CDFs, we provide an alternative representation via quantile functions, which forms the basis for its decomposition. Akin to the distances themselves, our proposed decompositions integrate over suitably assigned differences of the quantile functions and as such account for *any* distributional difference between F and G.

Section 6.2 introduces the novel decompositions together with extensive intuitive explanations. A particularly straightforward graphical illustration is available for $p = 1$, in which case the expressions in (6.2) and (6.3) coincide and are referred to as the *area validation metric* (AVM). The shift and dispersion terms in (6.1) then arise from simple comparisons of *central intervals* at *coverage levels* $\alpha \in [0, 1]$, i.e., intervals spanned by the $(1 \pm \alpha)/2$ quantiles of F and G. Roughly speaking, central intervals of differing lengths indicate differences in dispersion, while shifted intervals point to differing locations. The components in (6.1) are obtained by integrating over all coverage levels $\alpha \in [0, 1]$.

In Sections 6.3 and 6.4 we provide theoretical arguments that support the particular specifications of our decompositions.

Firstly, the decompositions behave naturally in settings where the distributions F and G are linked through additive shifts, symmetry relations, or a location-scale property. We also provide closed-form expressions for the components in the Gaussian case. Crucially, we prove that the proposed decompositions are *unique* in simultaneously satisfying a number of natural properties for (symmetric) location-scale families. This uniqueness is especially remarkable given that our decompositions operate directly on the quantile functions of the distributions.

Secondly, we show that the decompositions for the considered divergence measures mostly agree on which components are non-zero up to a subtle difference in the shift components. We further derive sensitivities of the divergences to differences in shift and dispersion. For symmetric distributions and with increasing power p, the p-Wasserstein distance exhibits an increasing sensitivity towards differences in dispersion. Furthermore, for Gaussian distributions, we show analytically that the Cramér distance weighs differences in dispersion even lower than the AVM (i.e., the Wasserstein distance WD_1 with the smallest power, $p = 1$).

Lastly, we derive comprehensive relations between our decompositions and suitable order relations of probability distributions. For each divergence function, there exist weak stochastic and dispersive orders such that the directed shift and dispersion components are (non-)zero if and only if the two distributions F and G are ordered accordingly. These properties further strengthen the theoretical backbone of our decompositions.

While extensive work has been done on decomposing proper scoring rules (Hersbach 2000; Bröcker 2009; Dimitriadis et al. 2021; Bracher et al. 2021a, among many others), the literature on decompositions of divergence functions into interpretable components is sparse. An exception is the exact decomposition of the 2-Wasserstein distance into the squared differences of the distributions' means and standard deviations, together with an analytically known remainder term capturing differences in shape (del Barrio et al., 1999; Irpino and Verde, 2015), which has recently been used in applications by Schefzik et al. (2021) and Lorenzo and Arroyo (2022). In contrast to this moment-based approach, which compares summary statistics that arise for each of the distributions separately, our decompositions are fully *nonparametric* in the sense of aggregating (integrating over) all distributional differences of F and G. Furthermore, our approach does not require a remainder term and is applicable to a range of divergence functions.

We note that our decompositions do not apply to other well-known divergence functions such as the Kullback–Leibler divergence or the Hellinger distance, as they lack a suitable representation in terms of quantile functions. Broadly speaking, these divergences are based on point-wise comparisons of probability density functions, without a notion of distance between the elements of $\mathbb{R}$ (the support of F and G), whereas the Wasserstein and Cramér distances consider *closeness* of F and G (Bellemare et al., 2017), i.e., the concentration of probability mass in nearby regions. This way of quantifying the distance between two distributions connects naturally to the notions of shifts and dispersion put forward in this work.

In Section 6.5, we illustrate the decompositions in two applications from the fields of climate science and economic survey design. Firstly, we take a closer look at the evaluation of climate models by Thorarinsdottir et al. (2020), who employ the Cramér distance to assess predictions of temperature extremes. For most models, our decompositions reveal systematic biases (upward for some, downward for others). Moreover, most models are

found to produce overdispersed predictive distributions. In a second application, we build upon work by Becker et al. (2023) who study how different histogram binning schemes impact responses in macroeconomic probabilistic surveys. Here, our decompositions serve to demonstrate that the submitted forecast distributions indeed change in ways which are coherent with the authors' pre-registered hypotheses.

The Appendix contains proofs and derivations together with additional illustrations, counterexamples and other details. Replication code is available at `https://github.com/resinj/replication_SD-Decomp`.

6.2. Quantile-based decompositions of divergence functions

We start by illustrating our decomposition in detail for the area validation metric (AVM) in Section 6.2.1. The more complex cases of the general p-Wasserstein distance (WD_p) and the Cramér distance (CD) will be addressed in Sections 6.2.2 and 6.2.3, respectively.

6.2.1. The area validation metric

For $p = 1$, the Wasserstein and l_p distances coincide and are referred to as the *area validation metric* (AVM),

$$\begin{aligned} \mathrm{AVM}(F, G) &= \int_{-\infty}^{\infty} \left|F(x) - G(x)\right| \mathrm{d}x \\ &= \int_0^1 \left|F^{-1}(\tau) - G^{-1}(\tau)\right| \mathrm{d}\tau. \end{aligned} \tag{6.4}$$

In the following we rely on the latter quantile-based representation, which we rewrite as

$$\begin{aligned} \mathrm{AVM}(F, G) &= \frac{1}{2} \int_0^1 \mathrm{AVM}_\alpha(F, G) \, \mathrm{d}\alpha, \quad \text{where} \\ \mathrm{AVM}_\alpha(F, G) &= \left|F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\alpha}{2}\right)\right| \\ &\quad + \left|F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right)\right|. \end{aligned} \tag{6.5}$$

Here, the integrand $\mathrm{AVM}_\alpha(F, G)$ simply compares the quantiles at levels $\frac{1-\alpha}{2}$ and $\frac{1+\alpha}{2}$, which span the *central intervals* with coverage probability α of the two distributions. The AVM results from integrating over all *coverage levels* $\alpha \in [0, 1]$.

The quantile-based and interval-based representations of the AVM are illustrated graphically in the top row of Figure 6.1. The left plot visualizes expression (6.4), while the right panel illustrates (6.5) in what we call a *quantile spread plot*. In a nutshell, the latter plot displays the (at the median) folded and (to coverage levels) rescaled quantile functions, i.e., the $(1 \pm \alpha)/2$ quantiles of F and G, which characterize the central intervals, as a function of the coverage $\alpha \in [0, 1]$. Twice the AVM then appears as the gray area between the two $\prec$-shaped curves if the dark gray area, which arises at coverages with disjoint central intervals, is counted twice.

In order to obtain four components as in (6.1), we use an α-wise decomposition of the integrand in (6.5),

$$\begin{aligned} \mathrm{AVM}_\alpha(F, G) = &\underbrace{\mathrm{Shift}^{\mathrm{AVM}}_{\alpha,+}(F, G)}_{\text{“}F\text{ shifted up”}} + \underbrace{\mathrm{Shift}^{\mathrm{AVM}}_{\alpha,-}(F, G)}_{\text{“}F\text{ shifted down”}} \\ &+ \underbrace{\mathrm{Disp}^{\mathrm{AVM}}_{\alpha,+}(F, G)}_{\text{“}F\text{ more dispersed”}} + \underbrace{\mathrm{Disp}^{\mathrm{AVM}}_{\alpha,-}(F, G)}_{\text{“}F\text{ less dispersed”}}. \end{aligned} \tag{6.6}$$

Using $[z]_+ := \max(z, 0)$ for the positive part of a real number $z \in \mathbb{R}$, we define the α-wise components

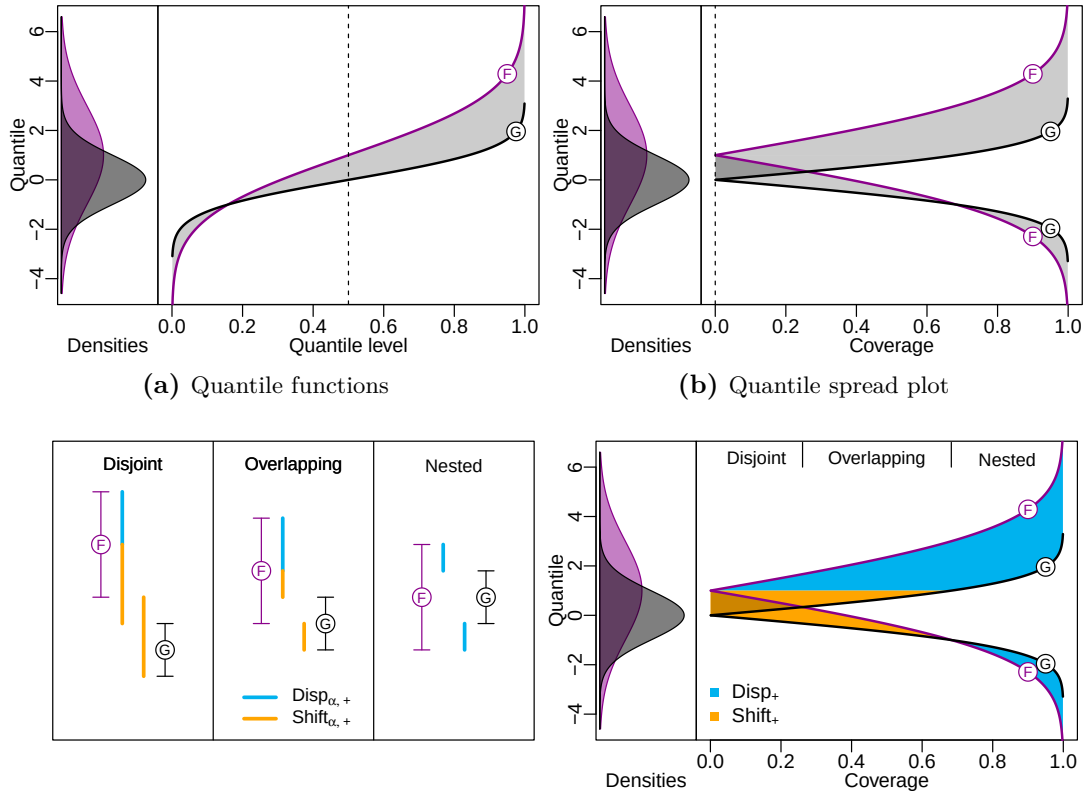
$$\begin{aligned} \mathrm{Shift}^{\mathrm{AVM}}_{\alpha,+}(F, G) := 2\Big[\min \{ &F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right), \\ &F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\alpha}{2}\right) \} \Big]_+ \end{aligned} \tag{6.7}$$

and

$$\begin{aligned} \mathrm{Disp}^{\mathrm{AVM}}_{\alpha,+}(F, G) := \Big[&\Big(F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right) \Big) \\ &- \Big(F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\alpha}{2}\right) \Big) \Big]_+ \end{aligned} \tag{6.8}$$

explained below. The remaining two α-wise components are symmetrically defined as

$$\begin{aligned} \mathrm{Shift}^{\mathrm{AVM}}_{\alpha,-}(F, G) &:= \mathrm{Shift}^{\mathrm{AVM}}_{\alpha,+}(G, F), \\ \mathrm{Disp}^{\mathrm{AVM}}_{\alpha,-}(F, G) &:= \mathrm{Disp}^{\mathrm{AVM}}_{\alpha,+}(G, F). \end{aligned} \tag{6.9}$$



(a) Quantile functions **(b)** Quantile spread plot

(c) AVM interval comparisons **(d)** AVM decomposition

Figure 6.1.: Illustration of the AVM decomposition for a pair of normal distributions, $F = \mathcal{N}(1, 2)$ and $G = \mathcal{N}(0, 1)$ together with their densities. The top row illustrates the AVM (gray shaded areas) in terms of the quantile-based formulation in (6.4) in panel (a) and the formulation in terms of central intervals in (6.5) in panel (b), where the shaded area equals $2 \times$ AVM. Panel (c) illustrates the AVM_α decomposition in (6.6) for three distinct configurations of the central intervals of two generic distributions F and G, which is discussed in the main text. Panel (d) shows the resulting decomposition of (twice) the AVM in the quantile spread plot across all coverage levels $\alpha \in [0, 1]$. The labels at the top of panel (d) indicate which configuration (as illustrated in panel (c)) occurs at each coverage level.

We henceforth refer to the two components with subscript '+' as the *plus* components as they quantify how F is, relative to G, shifted *upwards* and has an *increased* dispersion, respectively. Similarly, we refer to the terms with subscript '−' as the *minus* components.

The decomposition terms in (6.7)–(6.8) attribute the overall difference between the central α-interval endpoints of F and G—which is captured by the integrand $\mathrm{AVM}_\alpha(F, G)$ in (6.5)—to the components by more intricate interval comparisons. In a nutshell, the shift components capture twice the distance that one of the α-intervals needs to be moved to lie within the other, while the dispersion components measure by how much one needs to (de)compress the α-interval of G to have the same length as the α-interval of F.

The interval comparisons are illustrated in detail in the bottom left panel in Figure 6.1 for three distinct configurations of the central intervals of F and G (in purple and black, respectively). The two bars plotted between the intervals capture the two summands in (6.5). In the illustrations, the F-intervals are larger than the G-intervals by the blue parts, which are attributed to the dispersion component in (6.8). Note that (6.8) can be rewritten as the positive part of the difference between the lengths of the F- and G-intervals.

In the case of *nested* central intervals, the entire α-wise AVM in (6.5) is attributed to the dispersion component in (6.8). In the illustrated cases of *overlapping* or *disjoint* central intervals, the F-intervals lie higher than the G-intervals (in that the upper and lower endpoints are ordered in the same way). In this case, the shift component in (6.7) captures twice the minimum difference between the endpoints, which accounts for the remaining orange part of the colored bars.

If, on the other hand, the G-interval is wider than the F-interval, the difference in interval length is attributed to the minus dispersion component. Analogously, if the intervals are ordered differently, their difference in location is attributed to the minus shift component. On a technical level, these separations into plus and minus terms are achieved by the $[\cdot]_+$ operator in (6.7) and (6.8), respectively.

Notably, the α-wise dispersion components generalize upon the difference of the interquartile ranges (that arises for $\alpha = 0.5$ in (6.8)), and the shift components nest a comparison of the distributions' medians (for $\alpha = 0$ in (6.7)).

The bottom right panel of Figure 6.1 illustrates how the overall AVM decomposition arises through *integration* over all coverage levels $\alpha \in [0, 1]$. As in the top right panel, the entire colored area corresponds to twice the AVM. Notice that the dark yellow area

is counted twice, because the two bars overlap in disjoint interval configurations, as illustrated in panel (c).

Consequently, the components of the final decomposition as in (6.1) are defined as

$$\begin{aligned} \text{Shift}_{\pm}^{\text{AVM}}(F,G) &:= \frac{1}{2}\int_0^1 \text{Shift}_{\alpha,\pm}^{\text{AVM}}(F,G)\,\mathrm{d}\alpha, \\ \text{Disp}_{\pm}^{\text{AVM}}(F,G) &:= \frac{1}{2}\int_0^1 \text{Disp}_{\alpha,\pm}^{\text{AVM}}(F,G)\,\mathrm{d}\alpha, \end{aligned} \tag{6.10}$$

which yields the following result.

Proposition 6.2.1. *The* AVM *decomposition,*

$$\begin{aligned} \text{AVM}(F,G) = \text{Shift}_{+}^{\text{AVM}}(F,G) \;+\; \text{Shift}_{-}^{\text{AVM}}(F,G) \\ +\; \text{Disp}_{+}^{\text{AVM}}(F,G) \;+\; \text{Disp}_{-}^{\text{AVM}}(F,G), \end{aligned}$$

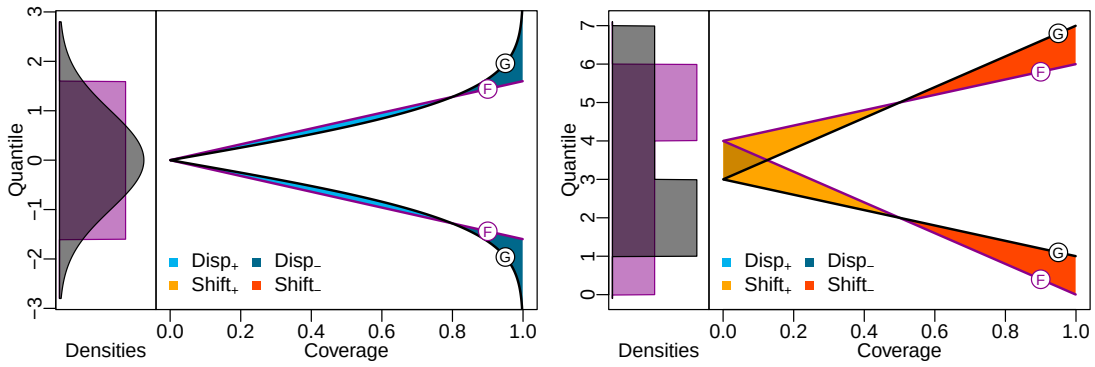
whose components are given in (6.7)–(6.10)*, is exact.*

Our decomposition is *nonparametric* by construction in the sense that it disaggregates the integrand $\text{AVM}_\alpha(F,G)$ in (6.5) at the fundamental quantile (or central interval) level and aggregates the resulting distributional differences through integration. Hence, the decomposition terms inherently capture all distributional discrepancies between F and G, as opposed to e.g., a moment-based decomposition into differences in means and variances as in del Barrio et al. (1999) and Irpino and Verde (2015).

At the α-wise level, at most one shift and one dispersion term in (6.7)–(6.9) can be positive. In contrast, all four components can be positive in the aggregated (integrated over all α levels) decomposition in Proposition 6.2.1, which we consider to be a natural feature of a nonparametric decomposition, as illustrated in the following examples.

Example 6.2.2. Here, we present two simple examples that lead to an AVM decomposition where both the plus and the minus shift or dispersion components are nonzero.

(a) The left panel of Figure 6.2 compares a uniform distribution $F = \mathcal{U}(-1.6, 1.6)$ with a standard normal distribution $G = \mathcal{N}(0,1)$. This example illustrates two distributions whose differences in central interval width vary for different coverage levels α, which yields nonzero plus and minus dispersion components, $\text{AVM}(F,G) =$



(a) AVM decomposition with nonzero dispersion terms **(b)** AVM decomposition with nonzero shift terms

Figure 6.2.: Graphical illustrations (as in Figure 6.1, panel (d)) of the AVM decompositions for the two distributional comparisons from Example 6.2.2 with densities on the left. Comparison (a) illustrated on the left leads to nonzero plus and minus dispersion components, while comparison (b) on the right leads to nonzero plus and minus dispersion components.

$\text{Disp}_+^{\text{AVM}}(F,G) + \text{Disp}_-^{\text{AVM}}(F,G) = 0.065 + 0.063$. As neither distribution exhibits a smaller dispersion at all coverage levels, having two positive dispersion components is a natural feature of our decomposition.

(b) The right panel of Figure 6.2 compares two mixtures of uniform distributions with a mirrored asymmetry, $F = 0.5 \times \mathcal{U}(0,4) + 0.5 \times \mathcal{U}(4,6)$, and $G = 0.5 \times \mathcal{U}(1,3) + 0.5 \times \mathcal{U}(3,7)$. While the width of all central intervals coincides (resulting in zero dispersion components), the locations of the central intervals are shifted in different directions for different coverage levels. This results in nonzero plus and minus shift components, $\text{AVM}(F,G) = \text{Shift}_+^{\text{AVM}}(F,G) + \text{Shift}_-^{\text{AVM}}(F,G) = 0.25 + 0.25$.

Further examples with (up to) four nonzero components are readily constructed. While such examples may occasionally be encountered in practice, the decomposition uncovers mostly clear differences in location and dispersion in our applications.

6.2.2. The p-Wasserstein distance

We next present a generalized decomposition for the p-th power of the p-Wasserstein distance in (6.2) for $p \in \mathbb{N}$. Here, we directly state the decomposition in its integral-form and dispense with a disaggregated α-wise treatment as in (6.6), which can be obtained by simply omitting the integrals in the following formulas.

Using the notation $z^{[p]} := \operatorname{sgn}(z) \cdot |z|^p$ as a shorthand for the *signed p-th power* of a real number $z \in \mathbb{R}$, we generalize the components given by (6.7)–(6.10) to

$$\begin{aligned} \operatorname{Shift}_+^{\mathrm{WD}_p}(F,G) := \frac{1}{2}\int_0^1 2\Bigg[\min\Bigg\{&\left(F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right)\right)^{[p]}, \\ &\left(F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\alpha}{2}\right)\right)^{[p]}\Bigg\}\Bigg]_+ \mathrm{d}\alpha, \end{aligned} \tag{6.11}$$

$$\begin{aligned} \operatorname{Disp}_+^{\mathrm{WD}_p}(F,G) := \frac{1}{2}\int_0^1 \Bigg[&\left(F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right)\right)^{[p]} \\ &- \left(F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\alpha}{2}\right)\right)^{[p]}\Bigg]_+ \mathrm{d}\alpha. \end{aligned} \tag{6.12}$$

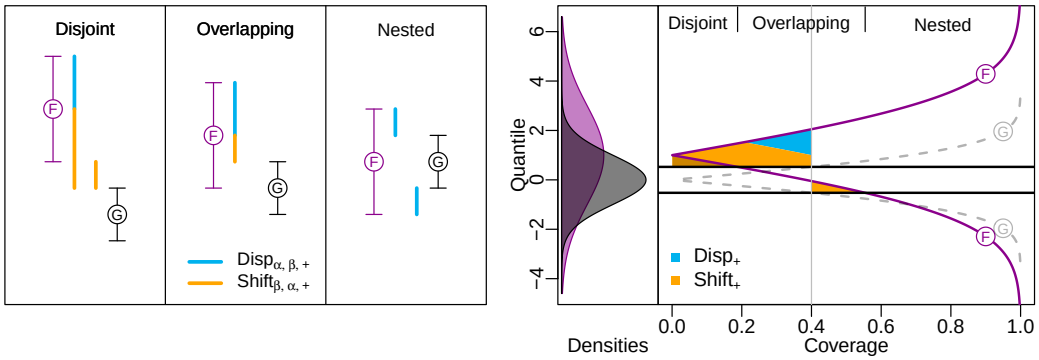
We again define $\operatorname{Shift}_-^{\mathrm{WD}_p}(F,G) := \operatorname{Shift}_+^{\mathrm{WD}_p}(G,F)$ and $\operatorname{Disp}_-^{\mathrm{WD}_p}(F,G) := \operatorname{Disp}_+^{\mathrm{WD}_p}(G,F)$ through symmetry and obtain the following result.

Proposition 6.2.3. *The* WD_p *decomposition,*

$$\mathrm{WD}_p(F,G) = \operatorname{Shift}_+^{\mathrm{WD}_p}(F,G) + \operatorname{Shift}_-^{\mathrm{WD}_p}(F,G) + \operatorname{Disp}_+^{\mathrm{WD}_p}(F,G) + \operatorname{Disp}_-^{\mathrm{WD}_p}(F,G),$$

whose components are given in (6.11)–(6.12)*, is exact.*

The interpretations of the respective components match the ones from Section 6.2.1 and Figure 6.1 when simply taking the signed p-th power of the distances between the respective interval ends considered in (6.11)–(6.12). For $p > 1$, taking the signed p-th power of interval end differences (opposed to $p = 1$ for the AVM) tends to result in a larger proportion of the Wasserstein distance being explained by a difference in dispersion than for the AVM. We study this phenomenon in mathematical detail in Section 6.3.2.



(a) CD interval comparisons **(b)** CD decomposition slice

Figure 6.3.: Illustration of the CD decomposition similar to Figure 6.1, panels (c) and (d). Panel (a) illustrates how the CD decomposition arises from individual comparisons of the central intervals of F and G for three distinct configurations. Panel (b) shows a quantile spread plot for the two normal distributions $F = \mathcal{N}(1, 2)$ and $G = \mathcal{N}(0, 1)$ together with their densities. The central interval of G at the fixed coverage level $\beta = 0.4$ is emphasized by the black horizontal lines. The corresponding *slice* of the Cramér distance is obtained by comparing the fixed central interval of G to all central intervals spanned by F. The overall CD and its decomposition are obtained by integrating all slices (i.e., integration across $\beta \in [0, 1]$). The labels at the top of panel (d) indicate which configuration (as illustrated in panel (a)) occurs at a given coverage level α (for the fixed level $\beta = 0.4$).

6.2.3. The Cramér distance

The *Cramér distance* (CD) or *integrated quadratic distance* corresponds to the squared l_2 metric,

$$\mathrm{CD}(F, G) = \int_{-\infty}^{\infty} \left|F(x) - G(x)\right|^2 \mathrm{d}x. \tag{6.13}$$

We start by providing a novel representation of the Cramér distance in terms of quantile functions, which is necessary to apply the ideas behind our quantile-based decompositions to the Cramér distance.

Proposition 6.2.4. *The Cramér distance in* (6.13) *can be expressed as*

$$\mathrm{CD}(F, G) = 2 \int_0^1 \int_0^1 \chi(\tau, \xi) \left|F^{-1}(\tau) - G^{-1}(\xi)\right| \mathrm{d}\tau \, \mathrm{d}\xi, \tag{6.14}$$

where

$$\chi(\tau,\xi) := \mathbb{1}\left\{\operatorname{sgn}(\tau-\xi) \neq \operatorname{sgn}\left(F^{-1}(\tau) - G^{-1}(\xi)\right)\right\}.$$

The indicator $\chi(\tau,\xi)$ used in (6.14) serves as an *incompatibility check* of the τ-quantile of F with the ξ-quantile of G: Whenever the order of the quantiles and quantile levels is at odds, i.e., $F^{-1}(\tau) > G^{-1}(\xi)$ despite $\tau < \xi$ or vice versa, the indicator function χ returns one and hence the pair of quantiles contributes to the Cramér distance.

Starting from the representation (6.14), a decomposition similar to the ones in Sections 6.2.1–6.2.2 arises that compares central intervals at differing coverage levels $\alpha, \beta \in [0,1]$ by setting

$$\begin{aligned}\operatorname{Shift}_+^{\mathrm{CD}}(F,G) := \frac{1}{2}\int_0^1\int_0^1 \Big[\min\Big\{ & F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\beta}{2}\right), \\ & F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\beta}{2}\right)\Big\}\Big]_+ \\ & + \Big[F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\beta}{2}\right)\Big]_+ \,\mathrm{d}\alpha\,\mathrm{d}\beta,\end{aligned} \tag{6.15}$$

$$\begin{aligned}\operatorname{Disp}_+^{\mathrm{CD}}(F,G) := \frac{1}{2}\int_0^1\int_0^\beta \Big[& \Big(F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\beta}{2}\right)\Big) \\ & - \Big(F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\beta}{2}\right)\Big)\Big]_+ \,\mathrm{d}\alpha\,\mathrm{d}\beta.\end{aligned} \tag{6.16}$$

We define the minus counterparts by $\operatorname{Shift}_-^{\mathrm{CD}}(F,G) := \operatorname{Shift}_+^{\mathrm{CD}}(G,F)$ and $\operatorname{Disp}_-^{\mathrm{CD}}(F,G) := \operatorname{Disp}_+^{\mathrm{CD}}(G,F)$, as before.

Proposition 6.2.5. *The* CD *decomposition,*

$$\mathrm{CD}(F,G) = \operatorname{Shift}_+^{\mathrm{CD}}(F,G) + \operatorname{Shift}_-^{\mathrm{CD}}(F,G) + \operatorname{Disp}_+^{\mathrm{CD}}(F,G) + \operatorname{Disp}_-^{\mathrm{CD}}(F,G),$$

whose components are given in (6.15)–(6.16)*, is exact.*

In analogy to the AVM decomposition in (6.7)–(6.10), the CD decomposition in (6.15)–(6.16) aggregates suitable comparisons of the central intervals via integration. Hence, we focus our discussion on the differences between the two decompositions. Most strikingly and in contrast to the previous decompositions, the components in (6.15)–(6.16) compare central intervals at differing coverage levels α and β through the double

integrals. For individual pairs of central intervals, differences to the AVM decompositions can be observed in the left-hand plot of Figure 6.3 that graphically illustrates the CD decomposition akin to the bottom-left plot in Figure 6.1.

While the integrand in the shift components in (6.15) resembles (6.7) in capturing the distance that one of the intervals needs to be moved to lie within the other, the factor of two is missing in (6.15). Moreover, in the case of disjoint intervals, the term in the last line of (6.15) yields an additional contribution capturing the distance that one of the intervals needs to be moved to overlap with the other.

The integrand in the dispersion components in (6.16) captures by how much one needs to (de)compress the β-interval of G to have the same length as the α-interval of F *if the coverage levels and interval lengths are at odds*. For example, if the α-interval of F is larger than the β-interval of G, different interval lengths indicate a difference in distributions only if $\alpha \leq \beta$, which is reflected by the integration boundary of the inner integral in (6.16). Otherwise, an increase in coverage naturally leads to an increased interval length even for identical distributions. In contrast, a shift between central intervals always hints at a distributional difference, regardless of coverage, as the central intervals of identical distributions are always nested.

The final components in (6.15)–(6.16) arise by integrating over the coverage levels α and β. As a joint graphical illustration of the double integral over α and β is challenging, we exemplarily fix the level of $\beta = 0.4$ in the right-hand plot of Figure 6.3 and illustrate the contributions in the integral over α for two normal distributions. In the Appendix, Figure D.1 shows equivalent plots for other values of β.

In the figure, we illustrate the comparison of the central intervals of F at all coverage levels α to the fixed central interval of G with coverage β, which is emphasized by the horizontal black lines. Contributions to the plus dispersion component that quantify by how much the central intervals of F are wider than the fixed central interval of G only arise for coverages $\alpha \leq \beta = 0.4$. No contributions to the plus shift component arise for coverages $\alpha \gtrsim 0.554$, as the 0.4-interval of G is strictly nested in the central intervals of F and hence no shifts between intervals arise. For coverages $0.188 \lesssim \alpha \lesssim 0.554$, the respective intervals are overlapping and the height of the orange area corresponds to the shift distance. We plot the area between the central intervals in such a way that it

is conveniently bounded by lower or upper interval ends, which leads to the break in the orange area at coverage $\alpha = 0.4$. Finally, for $\alpha \lesssim 0.188$, the intervals are disjoint, and the additional contribution in the last line of (6.15) results in the dark yellow area contributing twice.

As an aside, we note that the CD decomposition gives rise to a decomposition of the continuous ranked probability score (CRPS), a popular scoring rule used to evaluate probabilistic forecasts (Gneiting and Raftery, 2007).

Remark 6.2.6. If $G = \delta_y$ is a Dirac distribution at y, the Cramér distance reduces to the CRPS,

$$\mathrm{CD}(F, G) = \mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} |F(x) - \mathbb{1}(x \geq y)|^2 \mathrm{d}x.$$

As G has no variance in this case, one of the dispersion components vanishes, namely, $\mathrm{Disp}_{-}^{\mathrm{CD}}(F, G) = 0$. Denoting by m_F any median of F, the remaining components simplify to

$$\begin{aligned}
\mathrm{Disp}_{+}^{\mathrm{CD}}(F, G) &= \mathrm{Disp}^{\mathrm{CRPS}}(F, y) = \mathrm{CRPS}(F, m_F), \\
\mathrm{Shift}_{+}^{\mathrm{CD}}(F, G) &= \mathrm{Shift}_{+}^{\mathrm{CRPS}}(F, y) \\
&= \mathbb{1}(m_F > y) \times [\mathrm{CRPS}(F, y) - \mathrm{CRPS}(F, m_F)], \\
\mathrm{Shift}_{-}^{\mathrm{CD}}(F, G) &= \mathrm{Shift}_{-}^{\mathrm{CRPS}}(F, y) \\
&= \mathbb{1}(m_F < y) \times [\mathrm{CRPS}(F, y) - \mathrm{CRPS}(F, m_F)].
\end{aligned}$$

This decomposition is equivalent to the decomposition of the weighted interval score (an interval-based approximation of the CRPS) mentioned in Bracher et al. (2021a).

6.3. Theoretical properties of the decompositions

This section provides an in-depth analysis of the theoretical properties of the proposed decompositions. Section 6.3.1 establishes their natural behavior for distribution classes that are shifted, symmetric, of location-scale type and Gaussian. In Section 6.3.2, we contrast the sensitivity of the different divergence measures to distributional changes in dispersion and shift.

6.3.1. Basic properties

By construction, the components satisfy a simple symmetry.

Proposition 6.3.1 (Symmetry)**.** *For any $p \in \mathbb{N}$ and $\mathrm{D} \in \{\mathrm{AVM}, \mathrm{WD}_p, \mathrm{CD}\}$, we have*

$$\begin{aligned} \mathrm{Shift}_+^{\mathrm{D}}(F,G) &= \mathrm{Shift}_-^{\mathrm{D}}(G,F), \qquad \textit{and} \\ \mathrm{Disp}_+^{\mathrm{D}}(F,G) &= \mathrm{Disp}_-^{\mathrm{D}}(G,F). \end{aligned} \tag{6.17}$$

Hence, the properties that are outlined for the plus components (with subscript '+') also apply to the minus counterparts (with subscript '−') by symmetry.

Furthermore, it is easy to see that the dispersion components of the AVM and the CD are invariant to simple changes in location without imposing any distributional restrictions on F and G.

Proposition 6.3.2 (Dispersion invariant to shifts)**.** *For any distribution F and $s \in \mathbb{R}$, the shifted distribution F_s is given by $F_s(z) := F(z-s)$ for any $z \in \mathbb{R}$. Then, for all $s \in \mathbb{R}$ and $\mathrm{D} \in \{\mathrm{AVM}, \mathrm{CD}\}$, it holds that*

$$\mathrm{Disp}_\pm^{\mathrm{D}}(F_s, G) = \mathrm{Disp}_\pm^{\mathrm{D}}(F, G). \tag{6.18}$$

This invariance property of the dispersion components to simple location shifts is very natural. Unfortunately, it is not shared by the higher order Wasserstein distance decompositions with $p > 1$ as the simple shift by s does not cancel out in the dispersion components when subtracting signed p-th powers of the upper and lower interval end differences in (6.12).[1]

In order to analyze when one (or both) shift components vanish, we restrict attention to symmetric distributions. We call a distribution F symmetric if there exists a value $m \in \mathbb{R}$ such that $F^{-1}(\gamma) = 2m - F^{-1}(1-\gamma)$ for almost all $\gamma \in (0,1)$.[2] If there exists a unique median, then $m = F^{-1}(0.5)$, otherwise m is the midpoint of the interval of

[1] This also becomes obvious in the closed-form expressions for normal distributions in Appendix D.3 when F and G have the same mean but different variances and a shifted version F_s is compared to G.

[2] Note that admitting at most countably many points where the symmetry condition for the quantile function may be violated accounts for discontinuities in the quantile function. Such points form a null set in $[0,1]$ and can therefore be excluded from the integration domain without changing the value of an integral.

medians, which we henceforth call the *central median*. For symmetric distributions, the central median also coincides with the mean.

The following result shows that for two symmetric (but otherwise entirely flexible) distributions, at most one shift component is nonzero, and the direction of the shift agrees with the order of the medians.

Proposition 6.3.3 (Shift for symmetric distributions)**.** *Let F and G be symmetric distributions with central medians m_F and m_G and* $\mathrm{D} \in \{\mathrm{AVM}, \mathrm{WD}_p, \mathrm{CD}\}$*, $p \in \mathbb{N}$.*

(a) If $m_F \leq m_G$, then $\mathrm{Shift}_+^{\mathrm{D}}(F, G) = 0$.

(b) If the medians of F and G are unique, then

$$\mathrm{Shift}_+^{\mathrm{D}}(F, G) > 0 \quad \Longleftrightarrow \quad m_F > m_G. \tag{6.19}$$

The result in part (b) of Proposition 6.3.3 can be generalized to distributions with non-unique medians if the respective median intervals are non-nested. In this case, it still suffices to compare the central medians.

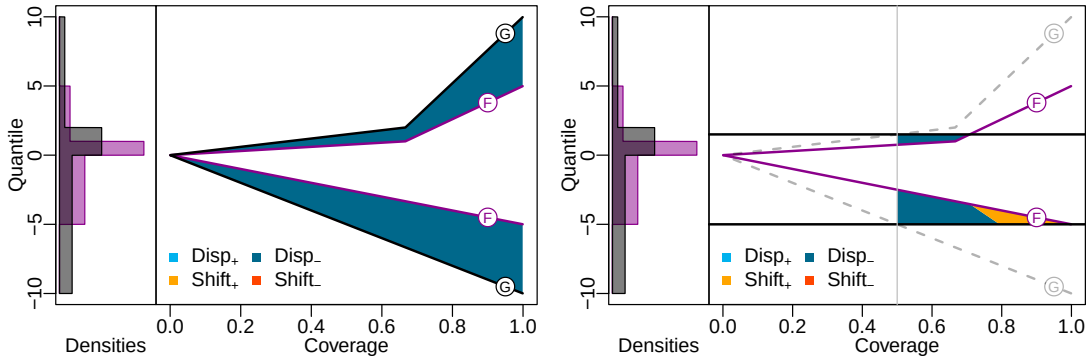
We continue to illustrate that our decompositions work as expected for distributional comparisons within location-scale families in that they reflect changes in location (scale) through a corresponding shift (dispersion) component. Notice that in the following, the location parameters ℓ_F and ℓ_G, and the scale parameters s_F and s_G are not necessarily means, medians or standard deviations of F and G, respectively.[3]

Proposition 6.3.4. *Let F and G be distributions from the same location-scale family, i.e., there exists a non-degenerate distribution H such that the quantile functions satisfy the relations $F^{-1} = s_F H^{-1} + \ell_F$ and $G^{-1} = s_G H^{-1} + \ell_G$ for some $\ell_F, \ell_G \in \mathbb{R}$ and $s_F, s_G > 0$.*

(a) For any $\mathrm{D} \in \{\mathrm{AVM}, \mathrm{WD}_p, \mathrm{CD}\}$ *with $p \in \mathbb{N}$, it holds that*

$$\mathrm{Disp}_+^{\mathrm{D}}(F, G) > 0 \quad \Longleftrightarrow \quad s_F > s_G. \tag{6.20}$$

[3]This is the case only if H is standardized to have mean or median zero, respectively, and variance equal to one.



(a) AVM decomposition **(b)** CD decomposition slice

Figure 6.4.: Illustration of the AVM and CD (for $\beta = 0.5$) decompositions for distributions F and G given in Example 6.3.5 from an asymmetric location-scale family where the distributions only differ in scale. See Figures 6.1 and 6.3 for detailed descriptions of the plots.

(b) The central medians of F and G are $m_F = s_F m_H + \ell_F$ and $m_G = s_G m_H + \ell_G$, respectively, where m_H denotes the central median of H. For any $p \in \mathbb{N}$, if $m_F \leq m_G$, then $\text{Shift}_+^{\text{WD}_p}(F, G) = 0$. Moreover, if the median of H is unique,

$$\text{Shift}_+^{\text{WD}_p}(F, G) > 0 \quad \iff \quad m_F > m_G. \tag{6.21}$$

Proposition 6.3.4 obviously also applies to pure location or pure scale families. As above, part (b) of Proposition 6.3.4 can be generalized to distributions with non-unique medians if the median intervals of F and G are non-nested by comparing the central medians in (6.21).

The following example shows that the second claim of Proposition 6.3.4 does not generalize to the CD decomposition. Nonetheless, an analogous equivalence holds for symmetric distributions by Proposition 6.3.3.

Example 6.3.5. Consider the asymmetric distributions $F = 0.5 \times \mathcal{U}[-5, 0] + \frac{1}{3} \times \mathcal{U}[0, 1] + \frac{1}{6} \times \mathcal{U}[1, 5]$ and $G = 0.5 \times \mathcal{U}[-10, 0] + \frac{1}{3} \times \mathcal{U}[0, 2] + \frac{1}{6} \times \mathcal{U}[2, 10]$, whose densities are shown in Figure 6.4. The distributions F and G stem from the same location-scale family as $H = F$ yields $F^{-1} = H^{-1}$ and $G^{-1} = 2F^{-1}$ with location parameters $m_F = m_G$. Hence, the right-hand side of the equivalence statement in Proposition 6.3.4 (b) does

not hold. However, as illustrated in the right-hand plot of Figure 6.4 (here, for fixed $\beta = 0.5$), the CD decomposition exhibits a nonzero shift component. In particular, it holds that $\mathrm{CD}(F,G) = \mathrm{Shift}_+^{\mathrm{CD}}(F,G) + \mathrm{Disp}_+^{\mathrm{CD}}(F,G) \approx 0.033 + 0.232$, while $\mathrm{AVM}(F,G) = \mathrm{Disp}_+^{\mathrm{AVM}}(F,G) = 23/12 \approx 1.917$.

The following theorem derives *uniqueness* of our decompositions if certain natural properties are warranted.

Theorem 6.3.6. *Let F and G be distributions from the same location-scale family with unique medians.*

(a) The shift-dispersion decomposition of $\mathrm{AVM}(F,G)$ *given in Proposition 6.2.1 is uniquely determined by the conditions* (6.17), (6.18), (6.20) *and* (6.21).

(b) If in addition F and G are symmetric, the shift-dispersion decomposition of $\mathrm{CD}(F,G)$ *given in Proposition 6.2.5 is uniquely determined by the conditions* (6.17), (6.18), (6.20) *and* (6.19).

Theorem 6.3.6 shows that our AVM and CD decompositions are unique amongst all possible decompositions in jointly satisfying the symmetry in (6.17), that dispersion terms are invariant to translations in (6.18), and that for (symmetric) location-scale families, non-negativity of the shift/dispersion terms agrees with the ordering of the location/scale terms in (6.19)–(6.21). As these properties are very natural to stipulate, Theorem 6.3.6 strongly supports the particular form of our decompositions. The uniqueness only applies to their integrated form, whereas the α-wise decomposition can of course be changed on (Lebesgue) null sets without affecting the resulting integral. A similar uniqueness condition cannot be established for the WD_p decomposition for $p > 1$ as the shift invariance of Proposition 6.3.2 cannot be invoked for $p > 1$.

We finally consider the case of normal distributions, which, arguably, form the most prominent location-scale family. Let $F = \mathcal{N}(\mu_F, \sigma_F^2)$ and $G = \mathcal{N}(\mu_G, \sigma_G^2)$ be normal distributions with means μ_F and μ_G and variances σ_F^2 and σ_G^2, respectively. In what follows, we use the shorthand notations $\widetilde{\mu} = |\mu_F - \mu_G|$, $\widetilde{\sigma} = |\sigma_F - \sigma_G|$ and $\widetilde{\rho} = \sqrt{\sigma_F^2 + \sigma_G^2}$. We further denote the standard normal density and distribution function by $\phi(\cdot)$ and $\Phi(\cdot)$, respectively. Here, we provide closed-form expressions for the decompositions of the

AVM and the CD. More involved formulas for the WD_p with arbitrary $p \in \mathbb{N}$ are given in Appendix D.3.

First, in the special case where $\widetilde{\sigma} = 0$, we get $\mathrm{AVM}(F, G) = \widetilde{\mu}$, which is entirely attributed to a single shift component. If $\widetilde{\sigma} \neq 0$, we get the closed-form expression

$$\mathrm{AVM}(F, G) = \widetilde{\mu}\big(2\Phi(\widetilde{\mu}/\widetilde{\sigma}) - 1\big) + 2\widetilde{\sigma}\phi(\widetilde{\mu}/\widetilde{\sigma}).$$

For its decomposition terms, if $\sigma_F > \sigma_G$, then

$$\mathrm{Disp}_{-}^{\mathrm{AVM}}(F, G) = 0 \quad \text{and} \quad \mathrm{Disp}_{+}^{\mathrm{AVM}}(F, G) = 2\widetilde{\sigma}\phi(0).$$

Furthermore (irrespective of whether $\sigma_F > \sigma_G$ holds), if $\mu_F > \mu_G$, then $\mathrm{Shift}_{-}^{\mathrm{AVM}}(F, G) = 0$ and

$$\mathrm{Shift}_{+}^{\mathrm{AVM}}(F, G) = \widetilde{\mu}\big(2\Phi(\widetilde{\mu}/\widetilde{\sigma}) - 1\big) + 2\widetilde{\sigma}\big(\phi(\widetilde{\mu}/\widetilde{\sigma}) - \phi(0)\big).$$

Hence, the shift and dispersion components agree, as expected, with the signs of differences in means and standard deviations, respectively.

For the CD of two normal distributions, we get the expression

$$\mathrm{CD}(F, G) = 2\widetilde{\rho}\phi(\widetilde{\mu}/\widetilde{\rho}) + \widetilde{\mu}\big(2\Phi(\widetilde{\mu}/\widetilde{\rho}) - 1\big) - \sqrt{2}\phi(0)\big(\sigma_F + \sigma_G\big).$$

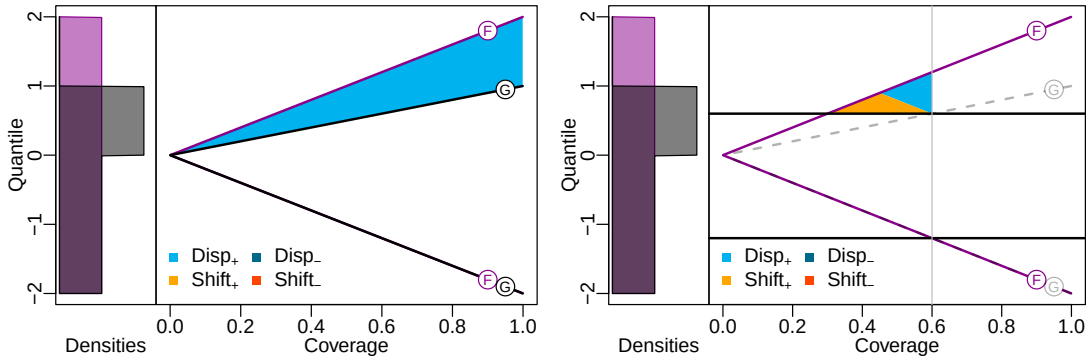
For its components, we obtain that if $\sigma_F \geq \sigma_G$, then $\mathrm{Disp}_{-}^{\mathrm{CD}}(F, G) = 0$ and

$$\mathrm{Disp}_{+}^{\mathrm{CD}}(F, G) = 2\widetilde{\rho}\phi(0) - \sqrt{2}\phi(0)(\sigma_F + \sigma_G).$$

Furthermore, if $\mu_F \geq \mu_G$, then $\mathrm{Shift}_{-}^{\mathrm{CD}}(F, G) = 0$ and

$$\mathrm{Shift}_{+}^{\mathrm{CD}}(F, G) = \widetilde{\mu}\big(1 - 2\Phi(-\widetilde{\mu}/\widetilde{\rho})\big) - 2\tau\big(\phi(0) - \phi(\widetilde{\mu}/\widetilde{\rho})\big).$$

Thus, the components of the CD also behave as expected for normal distributions.



(a) AVM decomposition **(b)** CD decomposition slice

Figure 6.5.: Illustration of the AVM and CD (for $\beta = 0.6$) decompositions for distributions F and G given in Example 6.3.9 for which only the CD exhibits a nonzero shift component. See Figures 6.1 and 6.3 for detailed descriptions of the plots.

6.3.2. Agreement and differences across distance measures

We now analyze how the decompositions of the considered divergence measures are related. We first show that they mostly agree on which components are nonzero. Subsequently, we use our decompositions to assess the relative importance of differences in location and dispersion of the analyzed distributions on the respective distance measures. Our results in the latter context shed light on fundamental differences between the distance measures under consideration.

Our first result shows that the decompositions of all considered divergences agree on which *dispersion* components are nonzero, without imposing any assumptions.

Proposition 6.3.7 (Positive dispersion)**.** *For any two distributions F and G, and $p \in \mathbb{N}$, it holds that*

$$\begin{aligned} \mathrm{Disp}_{\pm}^{\mathrm{WD}_p}(F,G) > 0 \quad &\iff \quad \mathrm{Disp}_{\pm}^{\mathrm{AVM}}(F,G) > 0 \\ &\iff \quad \mathrm{Disp}_{\pm}^{\mathrm{CD}}(F,G) > 0. \end{aligned}$$

Proposition 6.3.7 also implies that $\mathrm{Disp}_{\pm}^{\mathrm{WD}_p}(F,G) > 0 \iff \mathrm{Disp}_{\pm}^{\mathrm{WD}_q}(F,G) > 0$ for any $p \neq q$ by simply invoking the first equivalence for differing p and q.

A similar concordance property also holds for the shift components, however, without the equivalence with the CD components.

Proposition 6.3.8 (Positive shift)**.** *For any two distributions F and G, and $p \in \mathbb{N}$, it holds that*

$$\begin{aligned} \mathrm{Shift}_{\pm}^{\mathrm{WD}_p}(F,G) > 0 \quad &\Longleftrightarrow \quad \mathrm{Shift}_{\pm}^{\mathrm{AVM}}(F,G) > 0 \\ &\Longrightarrow \quad \mathrm{Shift}_{\pm}^{\mathrm{CD}}(F,G) > 0. \end{aligned} \tag{6.22}$$

In the following example, a positive shift component arises in the CD decomposition while the WD_p decomposition(s) include zero shift, which illustrates the missing equivalence between CD and AVM in Proposition 6.3.8.

Example 6.3.9. Consider the distributions $F = \mathcal{U}[-2,2]$ and $G = 0.5 \times \mathcal{U}[-2,0] + 0.5 \times \mathcal{U}[0,1]$, whose densities and folded quantile functions are shown in Figure 6.5. While the $\mathrm{AVM} = \mathrm{Disp}_{+}^{\mathrm{AVM}}(F,G) = 0.25$ is attributed entirely to an increase in dispersion of F relative to G, the $\mathrm{CD}(F,G) = \mathrm{Shift}_{+}^{\mathrm{CD}}(F,G) + \mathrm{Disp}_{+}^{\mathrm{CD}}(F,G) = 0.02083 + 0.02083$ is attributed to both shift and dispersion terms. (Notice for the later use of this example in Section 6.4.2 that the distribution F is strictly larger than G in the usual stochastic order.)

While the fact that the decompositions (almost) agree on which components are nonzero is conceptually reassuring, the nonzero components differ in (relative) magnitude. The following results shed light on the sensitivity of the considered divergence measures towards changes in shift and dispersion.

Theorem 6.3.10. *Let F and G be symmetric distributions, $F \neq G$, and $q > p \geq 1$ be positive integers. Then, it holds that*

$$\frac{\mathrm{Disp}_{+}^{\mathrm{WD}_p} + \mathrm{Disp}_{-}^{\mathrm{WD}_p}}{\mathrm{WD}_p(F,G)} \leq \frac{\mathrm{Disp}_{+}^{\mathrm{WD}_q} + \mathrm{Disp}_{-}^{\mathrm{WD}_q}}{\mathrm{WD}_q(F,G)}. \tag{6.23}$$

Theorem 6.3.10 shows that for symmetric distributions, the relative weight that the p-Wasserstein distance assigns to the dispersion components increases in its order p. Hence, Wasserstein distances of higher orders emphasize differences in dispersion as opposed to differences in location.

The symmetry condition in Theorem 6.3.10 is necessary to obtain a rigorous connection, as illustrated by Example 6.6.1 in Appendix D.5, which shows that inequality (6.23) is

not guaranteed to hold for asymmetric distributions, even when restricting attention to a location-scale family. Furthermore, Example 6.6.2 shows that the inequality also does not generally hold for unimodal distributions. Such examples are however rarely encountered in practice.

We now turn to a relative comparison of the AVM and the CD, which we formally establish for normal distributions using the closed-form expressions given at the end of Section 6.3.1.

Theorem 6.3.11. *Let $F = \mathcal{N}(\mu_F, \sigma_F^2)$ and $G = \mathcal{N}(\mu_G, \sigma_G^2)$ be normal distributions with $F \neq G$. Then,*

$$\frac{\mathrm{Disp}_+^{\mathrm{CD}} + \mathrm{Disp}_-^{\mathrm{CD}}}{\mathrm{CD}(F,G)} \leq \frac{\mathrm{Disp}_+^{\mathrm{AVM}} + \mathrm{Disp}_-^{\mathrm{AVM}}}{\mathrm{AVM}(F,G)}. \tag{6.24}$$

Theorem 6.3.11 establishes that for normal distributions, the AVM puts a higher emphasis on differences in dispersion in contrast to the CD. The relation in (6.24) is typically found in practice for other distributions as well, and counterexamples appear to be rare. Nonetheless, Example 6.6.3 shows that the inequality does not hold for arbitrary symmetric distributions (that are sufficient for the related statement in Theorem 6.3.10). We hypothesize that the following generalization of Theorem 6.3.11 to (possibly asymmetric) unimodal distributions may hold.

Conjecture 6.3.12. *Let F and G be unimodal distributions in the sense that the quantile functions are differentiable almost everywhere and the derivatives $(F^{-1})'$ and $(G^{-1})'$ are decreasing functions for $\alpha < \frac{1}{2}$ and increasing ones for $\alpha > \frac{1}{2}$. Then, it may hold that*

$$\frac{\mathrm{Disp}_+^{\mathrm{CD}} + \mathrm{Disp}_-^{\mathrm{CD}}}{\mathrm{CD}(F,G)} \leq \frac{\mathrm{Disp}_+^{\mathrm{AVM}} + \mathrm{Disp}_-^{\mathrm{AVM}}}{\mathrm{AVM}(F,G)}.$$

Such a result might appear counter-intuitive at first sight, as the shift components of the CD feature a factor of one-half that cancels out in the shift components of the AVM. The double integral and the capping of the CD dispersion component integrals at β apparently more than compensate for this halving.

Table 6.1.: Overview of the order relations used in this paper. All *strict* relations $F >_\bullet G$ are defined as $F \geq_\bullet G$ and $F \not\leq_\bullet G$. The unconventional (but equivalent; see (6.26)) definition of the usual stochastic order is used to highlight the similarity to the other stochastic order relations. The bracket "(cond.)" stands for conditions that are further discussed in Propositions 6.4.6 and 6.4.9.

Order name	Type	Symbol	Definition
Dispersive orders in Section 6.4.1:			
Dispersive	preorder	$F \geq_{\mathrm{D}} G$	$F^{-1}(\tau) - F^{-1}(\xi) \geq G^{-1}(\tau) - G^{-1}(\xi) \quad \forall\, 0 < \xi < \tau < 1$
Weak dispersive	preorder	$F \geq_{\mathrm{wD}} G$	$F^{-1}(\tau) - F^{-1}(1-\tau) \geq G^{-1}(\tau) - G^{-1}(1-\tau) \quad \forall\, 0.5 < \tau < 1$
Stochastic orders in Section 6.4.2:			
Usual stochastic	partial order	$F \geq_{\mathrm{S}} G$	$\min\left\{F^{-1}(\tau) - G^{-1}(\tau),\, F^{-1}(1-\tau) - G^{-1}(1-\tau)\right\} \geq 0$ $\forall\, 0.5 < \tau < 1$
Weak stochastic	preorder (cond.)	$F \geq_{\mathrm{wS}} G$	$\max\{F^{-1}(\tau) - G^{-1}(\xi),\, F^{-1}(1-\tau) - G^{-1}(1-\xi)\} \geq 0$ $\forall\, 0.5 < \tau, \xi < 1$
Relaxed stochastic	preorder (cond.)	$F \geq_{\mathrm{rS}} G$	$\max\{F^{-1}(\tau) - G^{-1}(\tau),\, F^{-1}(1-\tau) - G^{-1}(1-\tau)\} \geq 0$ $\forall\, 0.5 < \tau < 1$
Strong stochastic	strict partial order	$F >_{\mathrm{sS}} G$	$F \geq_{\mathrm{S}} G$ and $\exists \tau \in [0,1]$: $\min\{F(\tau) - G(\tau),\, F(1-\tau) - G(1-\tau)\} > 0$

6.4. Compatibility with stochastic order relations

In this section, we establish connections between our decomposition terms and some well-known stochastic orders (e.g., Müller and Stoyan, 2002; Shaked and Shanthikumar, 2007). We first show that our dispersion components align well with the dispersive order in Section 6.4.1. Subsequently, we relate our shift components to the usual stochastic order in Section 6.4.2. Table 6.1 provides an overview of all order relations used in this section. Here, we refer to orders that are related to the usual stochastic order as stochastic and use the term dispersive orders for stochastic variability orders. Throughout the section, we formulate the order conditions in terms of quantile levels τ and ξ for ease of exposition. Formulations in terms of coverage levels α and β that align more closely with the decomposition terms are obtained by replacing τ with $\frac{1+\alpha}{2}$ and ξ with $\frac{1+\beta}{2}$.

In what follows, we limit our analysis to probability distributions F and G with *continuous* quantile functions F^{-1} and G^{-1}, respectively. As alluded to at the end of Section 6.4.1, the results can be generalized to arbitrary distributions by slightly adapting the order relations to account for discontinuities in the quantile functions.

$$
\begin{array}{ccc}
F >_{\mathrm{D}} G & \Longrightarrow & \boxed{F >_{\mathrm{wD}} G \iff \left\{\begin{array}{l}\mathrm{Disp}_{-}^{\mathrm{D}}(F,G) = 0 \\ \mathrm{Disp}_{+}^{\mathrm{D}}(F,G) > 0\end{array}\right\}} \\
\Downarrow & & \Downarrow \\
F \geq_{\mathrm{D}} G & \Longrightarrow & \boxed{F \geq_{\mathrm{wD}} G \iff \mathrm{Disp}_{-}^{\mathrm{D}}(F,G) = 0}
\end{array}
$$

Figure 6.6.: Overview of the logical implications between the studied dispersive orders and the dispersion components of our decompositions.

6.4.1. Connections to the dispersive order

Here, we establish logical connections of (non-)zero dispersion components to order relations capturing differences in dispersion from the literature. The findings of this section are summarized in Figure 6.6.

The distribution F is said to be larger than G in *dispersive order*, $F \geq_{\mathrm{D}} G$, if for all $0 < \xi < \tau < 1$,

$$F^{-1}(\tau) - F^{-1}(\xi) \;\geq\; G^{-1}(\tau) - G^{-1}(\xi).$$

Note that the dispersive order is a preorder (which is reflexive and transitive) and not a partial order (which is reflexive, transitive and antisymmetric). As common for order relations, we define the *strict* relation $F >_{\mathrm{D}} G$ through $F \geq_{\mathrm{D}} G$ and $F \not\leq_{\mathrm{D}} G$, and equivalently for all other strict relations considered in this section.

The dispersive order turns out to be too strong for a logical equivalence with zero dispersion components to hold. To establish such an equivalence, we consider a weaker order relation. The distribution F is said to be larger than G in *weak dispersive order*, $F \geq_{\mathrm{wD}} G$, if for all $0.5 < \tau < 1$,

$$F^{-1}(\tau) - F^{-1}(1-\tau) \;\geq\; G^{-1}(\tau) - G^{-1}(1-\tau). \tag{6.25}$$

This order is also known under the name *quantile spread order* (Townsend and Colonius, 2005). It is easy to see from its definition that the weak dispersive order is a preorder. It gives rise to the following characterization result.

Theorem 6.4.1. *Let F and G be probability distributions with continuous quantile functions F^{-1} and G^{-1}, respectively. Then, for any $\mathrm{D} \in \{\mathrm{AVM}, \mathrm{WD}_p, \mathrm{CD}\}$, $p \in \mathbb{N}$, we have that*

(a) $F \geq_{\mathrm{wD}} G \iff \mathrm{Disp}_{-}^{\mathrm{D}}(F, G) = 0$;

(b) $F >_{\mathrm{wD}} G \iff \{\mathrm{Disp}_{-}^{\mathrm{D}}(F, G) = 0 \text{ and } \mathrm{Disp}_{+}^{\mathrm{D}}(F, G) > 0\}$.

Subject to the regularity condition of continuous quantile functions, which is further discussed below, part (a) of the theorem establishes an equivalence between a weak dispersive ordering of the distributions and a zero dispersion component in our decompositions. As shown in part (b), the equivalence extends naturally to an equivalence between a corresponding strict ordering and a unique nonzero dispersion component. Notably, the equivalences and implications in Theorem 6.4.1 hold for all considered divergence measures $\mathrm{D} \in \{\mathrm{AVM}, \mathrm{WD}_p, \mathrm{CD}\}$, $p \in \mathbb{N}$, by the equivalences from Proposition 6.3.7.

We continue to analyze the properties of the weak dispersive order. The following proposition shows that it is implied by the dispersive order.

Proposition 6.4.2. *Let F and G be probability distributions with continuous quantile functions F^{-1} and G^{-1}, respectively. Then,*

(a) $F \geq_{\mathrm{D}} G \implies F \geq_{\mathrm{wD}} G$;

(b) $F >_{\mathrm{D}} G \implies F >_{\mathrm{wD}} G$.

Notice that the implication in (b) is not trivial as if one order implies another, this does not necessarily mean that the same is true for the respective *strict orders*.

As summarized in Figure 6.6, the previous two results jointly show that a strict dispersive ordering implies a unique nonzero dispersion component in our decompositions. However, having a single nonzero dispersion term does not imply a dispersive ordering, even in its non-strict form. For example, the distributions F and G in Example 6.6.4 are not dispersively ordered despite the unique nonzero dispersion component.

Theorem 6.4.1 requires continuous quantile functions for F and G. The necessity of this condition is illustrated in the following example.

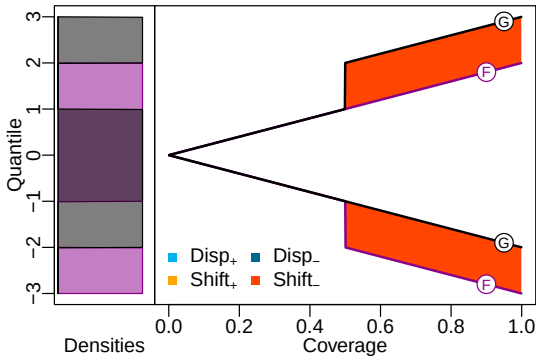


Figure 6.7.: Illustration of the AVM decomposition for distributions F and G given in Example 6.4.3 with discontinuous quantile functions such that F is strictly larger than G in weak dispersive order. See Figure 6.1 for a detailed description of the plot.

Example 6.4.3. Consider two mixtures of uniform distributions, $F = \frac{1}{4} \times \mathcal{U}[-3, -2] + \frac{3}{4} \times \mathcal{U}[-1, 2]$ and $G = \frac{3}{4} \times \mathcal{U}[-2, 1] + \frac{3}{4} \times \mathcal{U}[2, 3]$, with discontinuous quantile functions illustrated in Figure 6.7. The distribution F is strictly larger than G in weak dispersive order as the defining inequality is strict for $\tau = 0.75$ by left-continuity of the quantile functions, which results in a sort of asymmetric right-continuity in the lower part of the quantile spread plot. Despite the ordering, the decomposition does not produce a nonzero dispersion term, in clear contrast to part (b) of Theorem 6.4.1.

The issue encountered in Example 6.4.3 could be avoided by slightly adapting the definition of the weak dispersive order in (6.25) to require for any $0.5 \leq \tau < 1$ that

$$\sup\{x \mid F(x) \leq \tau\} - \inf\{x \mid F(x) \leq 1 - \tau\} \geq \sup\{x \mid G(x) \leq \tau\} - \inf\{x \mid G(x) \leq 1 - \tau\}.$$

This ensures that the lower and upper quantiles (as a function of the coverage) in the quantile spread plot are both right-continuous resulting in a symmetric directional continuity. We refrained from doing so in our analysis for ease of exposition and to align with the prevalent notion of the weak dispersive (or quantile spread) order.

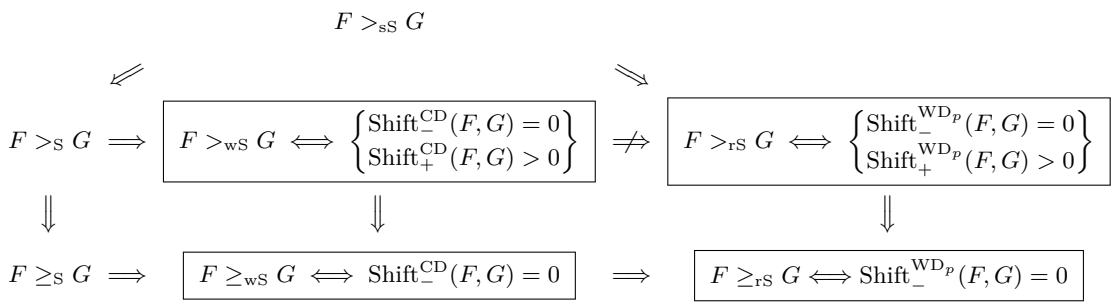


Figure 6.8.: Connections between the stochastic order relations introduced in Section 6.4.2 and the shift components of the CD and WD_p divergence measures (for any $p \in \mathbb{N}$).

6.4.2. Connections to the usual stochastic order

Establishing connections of (non)zero shift components to the usual stochastic order is more complicated than the connections to the dispersive orderings outlined in the previous Section 6.4.1. In a nutshell, zero shift components in the Cramér and Wasserstein decompositions give rise to a *weak* and a *relaxed* form of the stochastic order relation, respectively. Hence, we treat the Cramér and Wasserstein distances separately in the following two subsections. The results of this section are summarized in Figure 6.8.

The distribution F is said to be larger than G in the *usual stochastic order*, $F \geq_{\mathrm{S}} G$, if $F(x) \leq G(x)$ for all $x \in \mathbb{R}$, or, equivalently, if $F^{-1}(\tau) \geq G^{-1}(\tau)$ for all $0 < \tau < 1$. The usual stochastic order is a partial order, and its definition can be reformulated as

$$\min\left\{F^{-1}(\tau) - G^{-1}(\tau),\, F^{-1}(1-\tau) - G^{-1}(1-\tau)\right\} \;\geq\; 0 \tag{6.26}$$

for all $0.5 < \tau < 1$. The term used for this unconventional characterization arises in the shift components (see (6.7) with $\tau = \frac{1+\alpha}{2}$), thereby serving as a natural starting point to investigate the connection.

While $F \geq_{\mathrm{S}} G$ implies $\mathrm{Shift}_-^{\mathrm{D}}(F,G) = 0$ for all the considered distances $\mathrm{D} \in \{\mathrm{AVM}, \mathrm{WD}_p, \mathrm{CD}\}$, the usual stochastic order is too strong to establish an *equivalence* with zero shift components. Hence, the following two subsections present two relaxations of the usual stochastic order pertaining to the Cramér and Wasserstein distances, respectively.

A weak stochastic order based on the Cramér distance

To establish an equivalence relation with the shift components of the CD, we define the *weak stochastic order* in which the distribution F is larger than G, $F \geq_{\text{wS}} G$, if

$$\max\left\{F^{-1}(\tau) - G^{-1}(\xi),\, F^{-1}(1-\tau) - G^{-1}(1-\xi)\right\} \;\geq\; 0 \tag{6.27}$$

for all $0.5 < \tau, \xi < 1$. To the best of our knowledge, the weak stochastic order is a new order relation, and we discuss its properties after establishing its equivalence with the shift components of the CD.

Theorem 6.4.4. *Let F and G be probability distributions with continuous quantile functions F^{-1} and G^{-1}, respectively. Then,*

(a) $F \geq_{\text{wS}} G \iff \text{Shift}_{-}^{\text{CD}}(F, G) = 0$;

(b) $F >_{\text{wS}} G \iff \left\{\text{Shift}_{-}^{\text{CD}}(F, G) = 0 \;\; \textit{and} \;\; \text{Shift}_{+}^{\text{CD}}(F, G) > 0\right\}$.

The theorem establishes the equivalence of the (strict) weak stochastic order and (non)zero shift components of the CD, akin to Theorem 6.4.1.

The weak stochastic order is implied by the usual stochastic order, as detailed by the following proposition.

Proposition 6.4.5. *Let F and G be probability distributions with continuous quantile functions F^{-1} and G^{-1}, respectively. Then,*

(a) $F \geq_{\text{S}} G \implies F \geq_{\text{wS}} G$;

(b) $F >_{\text{S}} G \implies F >_{\text{wS}} G$.

As summarized in Figure 6.8, the previous results show that a strict stochastic ordering implies a unique nonzero shift component in the CD decomposition. On the other hand, having a single nonzero shift term in the CD does not imply a usual stochastic ordering, even in its non-strict form. For example, the distributions F and G in Example 6.6.4 are not stochastically ordered despite a unique nonzero shift component.

The weak stochastic order is a preorder for relatively broad classes of distributions.

Proposition 6.4.6. *The weak stochastic order is a preorder on sets of distributions with common support and continuous quantile functions.*

Example 6.6.4 illustrates the necessity of the common support assumption in Proposition 6.4.6.

A relaxed stochastic order based on Wasserstein distances

Similar to the treatment of the Cramér distance, we require a weaker form of a stochastic order relation to establish equivalence with nonzero shift components of the WD_p decompositions. To this end, we call F larger than G in *relaxed stochastic order*, $F \geq_{\mathrm{rS}} G$, if

$$\max\left\{F^{-1}(\tau) - G^{-1}(\tau),\, F^{-1}(1-\tau) - G^{-1}(1-\tau)\right\} \geq 0 \tag{6.28}$$

for all $0.5 < \tau < 1$. To the best of our knowledge, the relaxed stochastic order relation is new to the literature, and we briefly discuss its properties at the end of this section.

Theorem 6.4.7. *Let F and G be probability distributions with continuous quantile functions F^{-1} and G^{-1}, respectively. Then, for all $p \in \mathbb{N}$,*

(a) $F \geq_{\mathrm{rS}} G \iff \mathrm{Shift}_{-}^{\mathrm{WD}_p}(F, G) = 0$;

(b) $F >_{\mathrm{rS}} G \iff \left\{\mathrm{Shift}_{-}^{\mathrm{WD}_p}(F, G) = 0, \quad \mathrm{Shift}_{+}^{\mathrm{WD}_p}(F, G) > 0\right\}$.

This theorem establishes an equivalence between the (strict) relaxed stochastic order and (non)zero shift components of the WD_p shift components. Notably, the equivalences and implications in Theorem 6.4.7 hold for the Wasserstein distance of any order p by the equivalence from Proposition 6.3.8.

For the Wasserstein distances, a further complication arises because a *strict* stochastic ordering does not imply a nonzero shift component, as illustrated by Example 6.3.9. To establish a result akin to part (b) of Proposition 6.4.5, the strict stochastic order needs to be further strengthened. To this end, we define the *strong stochastic order*, $F >_{\mathrm{sS}} G$, as

$$\begin{gathered} F \geq_{\mathrm{S}} G \qquad \text{and} \qquad \exists \tau \in (0,1): \\ \min\{F^{-1}(\tau) - G^{-1}(\tau), F^{-1}(1-\tau) - G^{-1}(1-\tau)\} > 0. \end{gathered} \tag{6.29}$$

This relation is a *strict* partial order (irreflexive, antisymmetric and transitive) and implies the *strict* version of the usual stochastic order, which arises when replacing the minimum in (6.29) by a maximum.

Proposition 6.4.8. *Let F and G be probability distributions with continuous quantile functions F^{-1} and G^{-1}, respectively. Then,*

(a) $F \geq_{\mathrm{wS}} G \quad \Longrightarrow \quad F \geq_{\mathrm{rS}} G$;

(b) $F >_{\mathrm{sS}} G \quad \Longrightarrow \quad F >_{\mathrm{rS}} G$.

Part (a) of this proposition shows that the relaxed stochastic order is implied by the weak stochastic order, and hence, by invoking Proposition 6.4.5, also by the usual stochastic order. However, a corresponding implication fails to hold for their strict versions, as the distributions in Example 6.3.9 illustrate.

As summarized in Figure 6.8, part (b) of Proposition 6.4.8 together with Theorem 6.4.7 shows that a strong stochastic ordering implies a unique nonzero shift component in the WD_p decompositions.

We now establish that the relaxed stochastic order is a preorder under symmetry.

Proposition 6.4.9. *The relaxed stochastic order is a preorder on sets of symmetric distributions with continuous quantile functions.*

Example 6.6.5 shows that the common support assumption (that was imposed in Proposition 6.4.6) is not sufficient to establish transitivity for the *relaxed* stochastic order.

6.5. Applications

Here, we illustrate the decompositions in two applications from the fields of climate science and economic survey design.

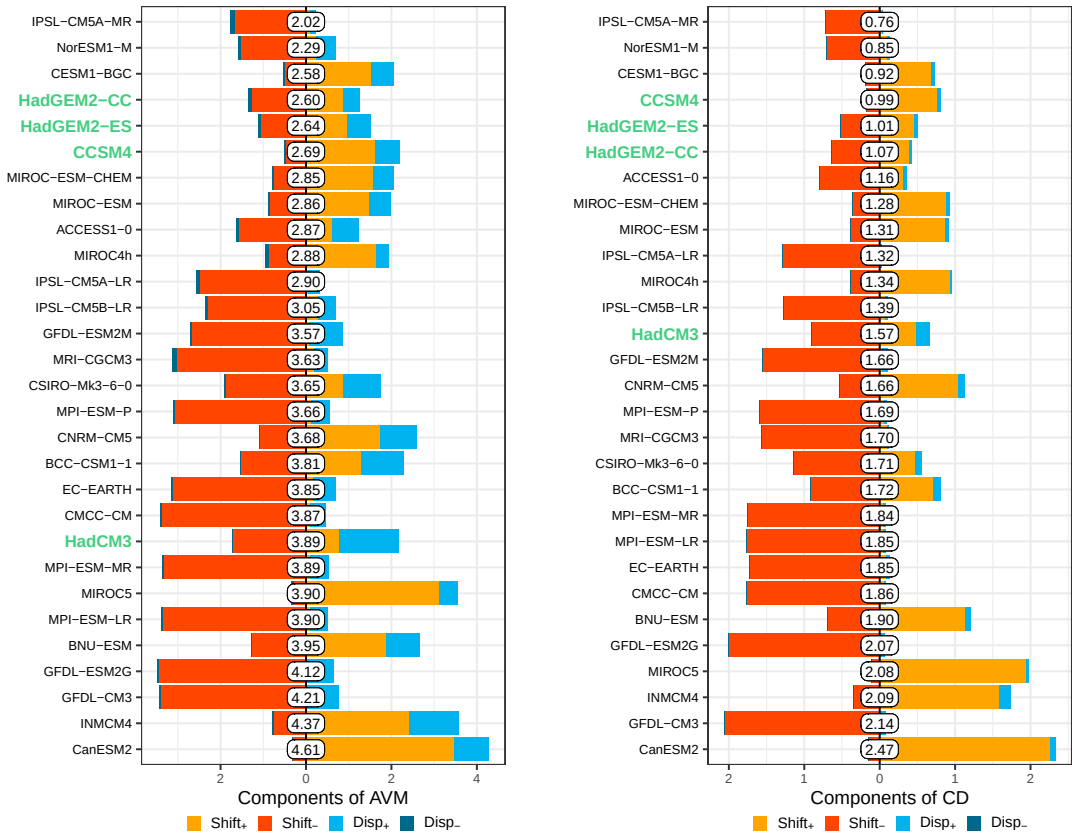
6.5.1. Prediction of seasonal temperature extrema

Our first application is from the field of climate modelling and illustrates how our decompositions can render forecast evaluations more interpretable. We revisit an evaluation of historical climate simulations from the Coupled Model Intercomparison Project (CMIP, Taylor et al. 2012) performed by Thorarinsdottir et al. (2020). Our focus is on monthly maximum temperatures (TXx) over Europe during the Boreal summer months June, July, and August.

Thorarinsdottir et al. (2020) compare the empirical distributions of these temperature extremes over the years 1979–2005 according to various data sources to corresponding forecast distributions from a variety of models. We replicate a comparison between the primary data source used in the paper (HadEX2, Donat et al. 2013) and 29 different forecasting models from the CMIP5 project (Sillmann et al., 2013). For 136 grid cells of size 3.75° (longitude) × 2.5° (latitude) covering European land masses, the cell-wise Cramér distances between the empirical and model-based distributions are computed and subsequently averaged (Figure 4 in Thorarinsdottir et al. 2020). This results in a ranking of the different models in terms of their capacity to predict the distribution of temperature extremes.

We rerun these computations[4], apply both the Cramér distance and the area validation metric and compute the respective decompositions. The results are displayed in Figure 6.9. Several relevant patterns emerge, which are not discernible based on the average divergences from Thorarinsdottir et al. (2020). Firstly, there is a clear dominance of one of the shift components for most models. This indicates that the differences between empirical and predicted distributions tend to be systematic across grid cells. An interesting exception is the model `HadGEM2-ES`, where both shift components are of similar size. As shown in Figure 6.10, this mixed pattern results from different shifts in

[4]The results for the average Cramér distance for the models MIROC5 and MIROC-ESM_CHEM reported here differ from the ones in Thorarinsdottir et al. (2020, Figure 4, left) for unknown reasons.

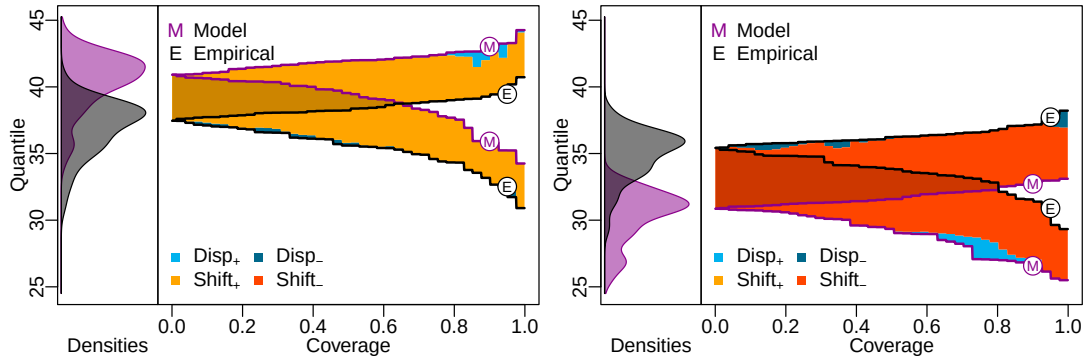


(a) AVM decompositions **(b)** CD decompositions

Figure 6.9.: Evaluation of forecasts for monthly temperature maxima (TXx) in summer months from 29 climate models (given in the rows), averaged over 136 grid cells in Europe, sorted by their average (across grid cells) AVM (left) and CD (right). We decompose these distances between the empirical distributions (HadEX2) and the model forecasts into our shift and dispersion components, whose magnitude is shown by the colored bars. To facilitate visual distinction, opposing shift and dispersion components are drawn in different directions. The models that are discussed in the text are highlighted in green.

different grid cells rather than simultaneously positive shift components within the same grid cell. Secondly, concerning the dispersion components, it can be seen that the model forecasts are consistently more dispersed than the empirical distributions.

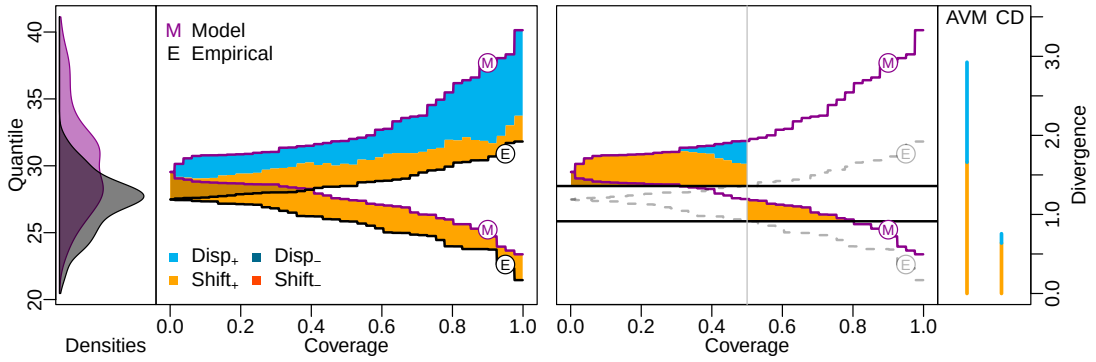
The application moreover illustrates that the AVM tends to give more weight to dispersion components than the CD, as discussed in Section 6.3.2. While for some models,



(a) AVM decomposition (one grid cell) **(b)** AVM decomposition (another grid cell)

Figure 6.10.: Two examples of comparisons on specific grid cells between the empirical TXx distributions (HadEX2, black) and the predictive distributions from the `HadGEM2-ES` forecasting model (purple). In each panel, we show kernel density estimates of the two distributions, as well as their AVM decompositions (based on empirical CDFs). In each of the grid cells, exactly one shift component is non-zero: The model overpredicts in the left, while underpredicting in the right plot. See Figure 6.1 for details on the graphical display of the AVM decomposition.

the dispersion component represents a substantial part of the overall AVM, the dispersion components are largely negligible for the CD. In Figure 6.11 we illustrate this behavior for a selected grid cell. The different relative importance of shift and dispersion components also explains some of the differences in model rankings across the two divergences. Most notably, the `HadCM3` model receives a large dispersion component under the AVM, leading to a considerably worse ranking than under the CD. Similarly, the `HadGEM2CC` and `CCSM4` models (ranks 4 and 6, respectively, under AVM) change places under the CD. The reason for these changes is that differences in dispersion, which play a relevant role in the AVM, become negligible in the CD.



(a) AVM decomposition **(b)** CD decomposition slice

Figure 6.11.: Comparison of the decompositions of the AVM and CD for a selected grid cell (HadEX2 data in black against the `HadGEM2-ES` model in purple). For the selected grid cell, the two distributions visibly differ in dispersion, which is picked up more strongly by the AVM (second panel) than the CD (third panel). Note that the CD decomposition is only shown for the level $\beta = 0.5$, but as can be seen from the fourth panel, the overall decomposition gives similarly little weight to Disp_+. See Figures 6.1 and 6.3 for details on the graphical displays of the decompositions.

6.5.2. Elicitation of inflation predictions

Our second application considers economic surveys on households' expectations about future inflation, which have become an important tool for central banks over the previous decades (Coibion et al., 2022). Inspired, among others, by the call of Manski (2004) for probabilistic forecasts, recent surveys such as the Survey of Consumer Expectations (SCE) of the Federal Reserve Bank of New York are often in a probabilistic form, where the respondents assign probabilities to pre-specified bins, resulting in histogram-like forecasts. The issued probabilities together with the right endpoints of the bins elicit precise information on the forecasted CDFs. However, this survey format has come under question as the respondents' elicited distributions are found to be sensitive to the bin specification (Schwarz et al., 1985; Becker et al., 2023). This effect is especially disconcerting in times of varying inflation rates that necessitate adjustments in the bin specifications.

Becker et al. (2023) focus on the undesirable effect that a given binning specification has on the responses in a preregistered experimental study on the platform Prolific (`www.prolific.co`), where randomly selected respondents obtain varying bin specifications

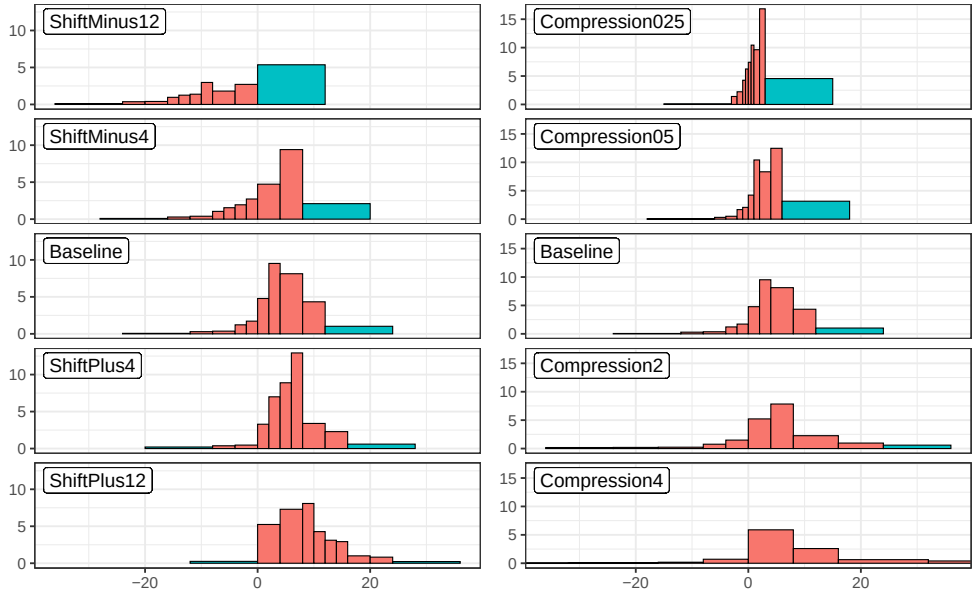


Figure 6.12.: Density (histogram) forecasts averaged over all survey respondents for the shift and compression treatments of Becker et al. (2023). All treatments include an open interval on the left and right in blue, which we truncate such that it has a width of 12 percentage points purely for illustrative purposes.

as a treatment when issuing their probabilistic expectations. The baseline specification follows the traditional binning used in the SCE. The authors consider (among others) "shift" and "compression" treatments, where the bins are shifted to the right or left, and compressed or decompressed, as illustrated in Figure 6.12.

While it would be desirable that responses are unaffected by the given binning, Becker et al. (2023) find significant shift and compression effects in the average responses that are closely related to the implemented changes in the binning. However, their methodology is limited to analyzing the means and standard deviations that are obtained from fitting a parametric beta distribution to the individual responses (Engelberg et al., 2009), and to the probability of the binarized event of deflation.

However, given that the survey is probabilistic in nature, so should be its evaluation. Hence, we refine the results of Becker et al. (2023) by considering our decompositions of the AVM and CD of the aggregated response distributions. Our decompositions in shift and dispersion components are particularly suitable for the shift and compression

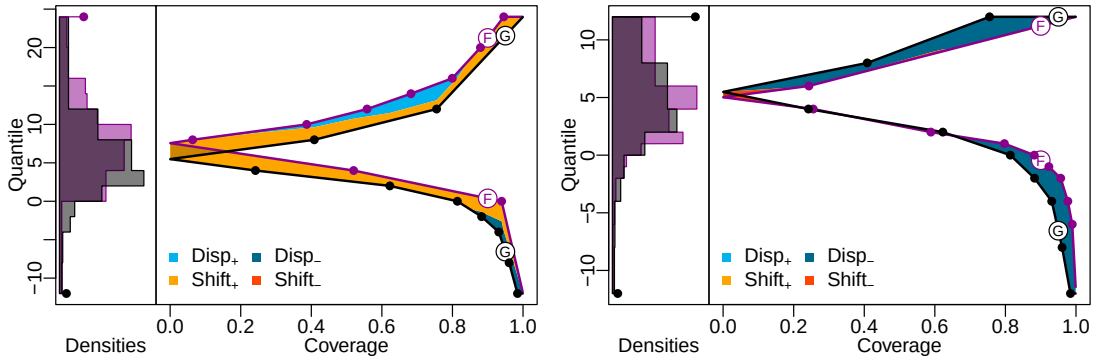
Table 6.2.: Approximate Cramér distances and area validation metrics together with their (approximated) decompositions of the average histogram forecast under the various treatments (F) described in the first column in comparison to the average forecast under the Baseline treatment (G).

	Cramér Distance (CD)					Area Validation Metric (AVM)				
Treatment (F)	CD	Shift_+	Shift_-	Disp_+	Disp_-	AVM	Shift_+	Shift_-	Disp_+	Disp_-
ShiftMinus12	0.820	0.000	0.534	0.286	0.000	4.556	0.000	1.823	2.733	0.000
ShiftMinus4	0.081	0.000	0.070	0.011	0.000	1.222	0.000	0.743	0.443	0.037
ShiftPlus4	0.068	0.060	0.000	0.001	0.007	0.996	0.649	0.000	0.118	0.228
ShiftPlus12	0.206	0.198	0.000	0.008	0.001	2.025	1.626	0.000	0.303	0.096
Compression025	0.123	0.000	0.108	0.001	0.015	1.344	0.000	0.842	0.051	0.451
Compression05	0.050	0.000	0.017	0.000	0.033	0.859	0.000	0.079	0.000	0.780
Compression2	0.149	0.007	0.000	0.142	0.000	2.442	0.097	0.000	2.345	0.000
Compression4	0.488	0.120	0.000	0.368	0.000	5.060	0.358	0.000	4.702	0.000

treatments of Becker et al. (2023). Table 6.2 shows the AVM and CD together with the four components of the decompositions comparing the different treatments (F) to the baseline distribution (G). As the distributions shown in Figure 6.12 identify the respective CDFs only at the values separating the histogram bins, we use the approximation of the decomposition terms described in Appendix D.9 and shown in Figure 6.13 for two exemplary treatments. The solid points in the quantile spread plots show the known quantiles and a linear interpolation is used in between. In the tails, a conservative extrapolation is used that we describe in detail in Appendix D.9.

For all four shift treatments, Table 6.2 shows a large shift component in the anticipated direction for both, the AVM and the CD. As already noted by Becker et al. (2023), the "artificial truncation" of the bins (especially in ShiftMinus12) that can be observed in Figure 6.12 together with the recent surge in inflation rates results in an increase in dispersion. As expected from the theoretical results of Section 6.3.2, the relative magnitude (among the entire divergence) of the shift components is smaller for the AVM than for the CD.

In the compression treatment, we find equivalent effects in the dispersion components in the expected direction. The compression treatments with factors 0.25 and 0.5 entail—perhaps surprisingly—large shifts, which can be traced back to the artificial truncation combined with the high inflation rates at the time of the survey in December 2021.



(a) AVM decomposition for ShiftPlus12 **(b)** AVM decomposition for Compression05

Figure 6.13.: Illustrations of the approximate AVM decompositions for the ShiftPlus12 (left) and Compression05 (right) treatments (F) against the baseline treatment (G). Point masses are represented by solid dots in the density plots. Known quantiles are marked with solid dots in the quantile spread plots. See Figure 6.1 for a detailed description of the plots.

The results in this section supplement the parametric analysis of Becker et al. (2023) by adding a nonparametric verification that takes into account the full distributional differences between the baseline and the treated probability forecasts.

6.6. Discussion

In the present paper, we introduce decompositions into dispersion and shift components for various statistical distances. The decompositions meaningfully attribute the overall distance to differences in location and variability. They behave as expected in clear-cut special cases such as pairs of symmetric distributions and distributions from the same location-scale family, while being applicable to arbitrary pairs of distributions. Furthermore, the decompositions shed light on the sensitivities of the studied divergences towards differences in location and dispersion. The decompositions are compatible with the usual stochastic and dispersive order relations, and we establish correspondences to suitably constructed weakened order relationships. Finally, we demonstrate the practical use of the decompositions in two case studies.

Both our theoretical results and case studies indicate that the Cramér distance is rather insensitive to differences in dispersion, and puts a strong focus on differences in location.

The area validation metric (and higher-order Wasserstein distances) on the other hand emphasize differences in dispersion. Besides these natural differences, we note that the Cramér distance has an important advantage when used for forecast evaluation as in the application to climate predictions. Unlike the area validation metric (and higher-order Wasserstein distances), it is a *proper divergence metric* (Thorarinsdottir et al., 2013), which rewards truthful predictions.

As a consequence of our interval-based approach, the proposed decompositions attribute the distance between two distributions F and G entirely to differences in location and dispersion. Of course, distributions are often characterized by additional (higher-order) properties such as skewness or kurtosis attributed to their shape. In our decompositions, differences in higher-order properties sometimes lead to both shift or dispersion components being nonzero simultaneously (e.g., in Example 6.2.2), but there are no clear-cut connections. The *nonparametric* nature of our decompositions that aggregate fundamental comparisons of central intervals does not attest to differences in shape, while simple decompositions based on comparing summary statistics, which may do so, provide a rather superficial comparison and are not available for most of the studied distances. Accommodating additional shape components likely requires considerably revised techniques and is reserved for future work.

Finally, the decompositions are thus far descriptive tools that appear to be promising both from a theoretical and applied perspective. A next step will be to develop inference techniques that assess the statistical significance of the components in standard one- and two-sample settings as well as more involved settings such as aggregated inflation expectations. While resampling techniques such as permutation tests or the bootstrap are well suited to assess the statistical significance of the overall distance, assessing the significance of individual components is challenging. As the null hypotheses of no difference in shift or location are complex composite hypotheses, simulating them does not seem feasible. A careful look at the sum of the dispersion components of the area validation metric reveals that it can be rewritten as a sum of conditional expectations, which may well facilitate inference for this particular component. We view such developments as a natural avenue for future research.

Appendix D

The Appendix contains proofs and derivations together with additional illustrations, counterexamples, and details on approximations. Section D.1 graphically illustrates the CD decomposition for a range of β values. We provide closed-form expressions for the WD_p decomposition in Section D.2. Section D.4 provides additional counterexamples. Section D.7 provides a general purpose approximation of the CD decomposition and gives details on the approximations used in Section 6.5.2. Finally, Sections D.10–D.15 contain all proofs of the results presented in Sections 6.2–6.4, respectively.

D.1. Graphical illustration of the CD decomposition

Figure D.1 contains graphical illustrations as in Figure 6.3, panel (b), at various levels $\beta \in \{0.1, 0.2, \ldots, 0.9\}$.

D.2. Closed-form expressions for normal distributions

We supplement the closed-form expressions given in the main text for the area validation metric and the Cramér distance between two normal distributions with formulas for the Wasserstein distance of order p.

D.3. The WD_p decomposition for normal distributions

Here, we present closed-form expressions for the decomposition terms of the p-th power of the p-Wasserstein distance with $p \in \mathbb{N}$ for two normal distributions, $F = \mathcal{N}(\mu_F, \sigma_F^2)$ and $G = \mathcal{N}(\mu_G, \sigma_G^2)$. As in Section 6.3, we use of the shorthand notations $\widetilde{\mu} = |\mu_F - \mu_G|$ and $\widetilde{\sigma} = |\sigma_F - \sigma_G|$.

The following formulas are based on expressions $m_p(\mu, \sigma, a)$ for the p-th moments of a normal distribution $\mathcal{N}(\mu, \sigma^2)$ that is *truncated* from below at the value $a \in \mathbb{R}$. Orjebin (2014) provides the recursive formula

$$m_p(\mu, \sigma, a) = (p-1)\sigma^2 m_{p-2}(\mu, \sigma, a) + \mu m_{p-1}(\mu, \sigma, a) + \sigma \frac{a^{p-1}\phi(\frac{a-\mu}{\sigma})}{1 - \Phi(\frac{a-\mu}{\sigma})}$$

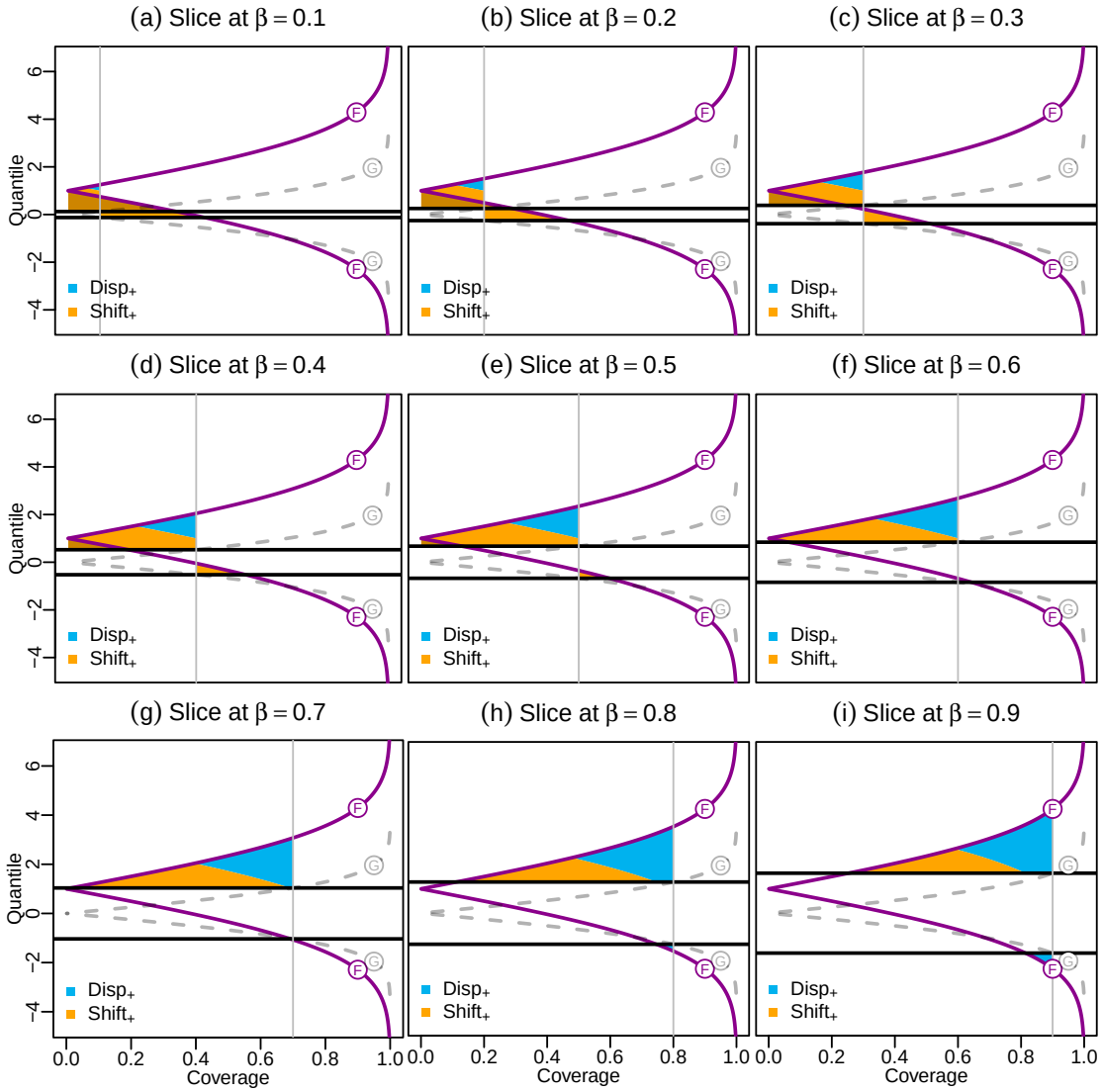


Figure D.1.: Graphical illustrations as in Figure 6.3 at various $\beta \in \{0.1, 0.2, \ldots, 0.9\}$ levels.

for $\mu \in \mathbb{R}, \sigma > 0$ and $p \in \mathbb{N}$ with $m_0(\mu, \sigma, a) = 1$ and $m_{-1}(\mu, \sigma, a) = 0$. Then, the Wasserstein distance between two normal distributions is given by

$$\mathrm{WD}_p(F, G) = \begin{cases} \widetilde{\mu}^p & \text{if} \quad \sigma_F = \sigma_G, \\ \Phi(\widetilde{\mu}/\widetilde{\sigma}) m_p(\widetilde{\mu}, \widetilde{\sigma}, 0) + \Phi(-\widetilde{\mu}/\widetilde{\sigma}) m_p(-\widetilde{\mu}, \widetilde{\sigma}, 0) & \text{if} \quad \sigma_F \neq \sigma_G. \end{cases}$$

If the variances are equal, the p-Wasserstein distance is assigned to a single shift component,

$$\sigma_F = \sigma_G \quad \Longrightarrow \quad \mathrm{WD}_p(F,G) = \widetilde{\mu}^p = \begin{cases} \mathrm{Shift}_+^{\mathrm{WD}_p}(F,G), & \text{if } \mu_F > \mu_G, \\ \mathrm{Shift}_-^{\mathrm{WD}_p}(F,G), & \text{if } \mu_F < \mu_G, \end{cases}$$

and the dispersion components are zero (as well as the opposing shift component).

In contrast, for differing variances, we get

$$\sigma_F > \sigma_G \quad \Longrightarrow \quad \begin{cases} \mathrm{Disp}_+^{\mathrm{WD}_p}(F,G) & = m_p(\widetilde{\mu},\widetilde{\sigma},\widetilde{\mu}) + (1-\Phi(\widetilde{\mu}/\widetilde{\sigma}))m_p(-\widetilde{\mu},\widetilde{\sigma},0) \\ & \quad -\Phi(\widetilde{\mu}/\widetilde{\sigma})m_p(\widetilde{\mu},\widetilde{\sigma},0), \\ \mathrm{Disp}_-^{\mathrm{WD}_p}(F,G) & = 0, \end{cases}$$

and the above dispersion terms are swapped if $\sigma_F < \sigma_G$.

Finally, given differing variances and an ordering of the means, we get

$$\mu_F \geq \mu_G \quad \Longrightarrow \quad \begin{cases} \mathrm{Shift}_+^{\mathrm{WD}_p}(F,G) = 2\Phi(\widetilde{\mu}/\widetilde{\sigma})m_p(\widetilde{\mu},\widetilde{\sigma},0) - m_p(\widetilde{\mu},\widetilde{\sigma},\widetilde{\mu}) \\ \mathrm{Shift}_-^{\mathrm{WD}_p}(F,G) = 0 \end{cases}$$

and the above shift terms are swapped if $\mu_F < \mu_G$.

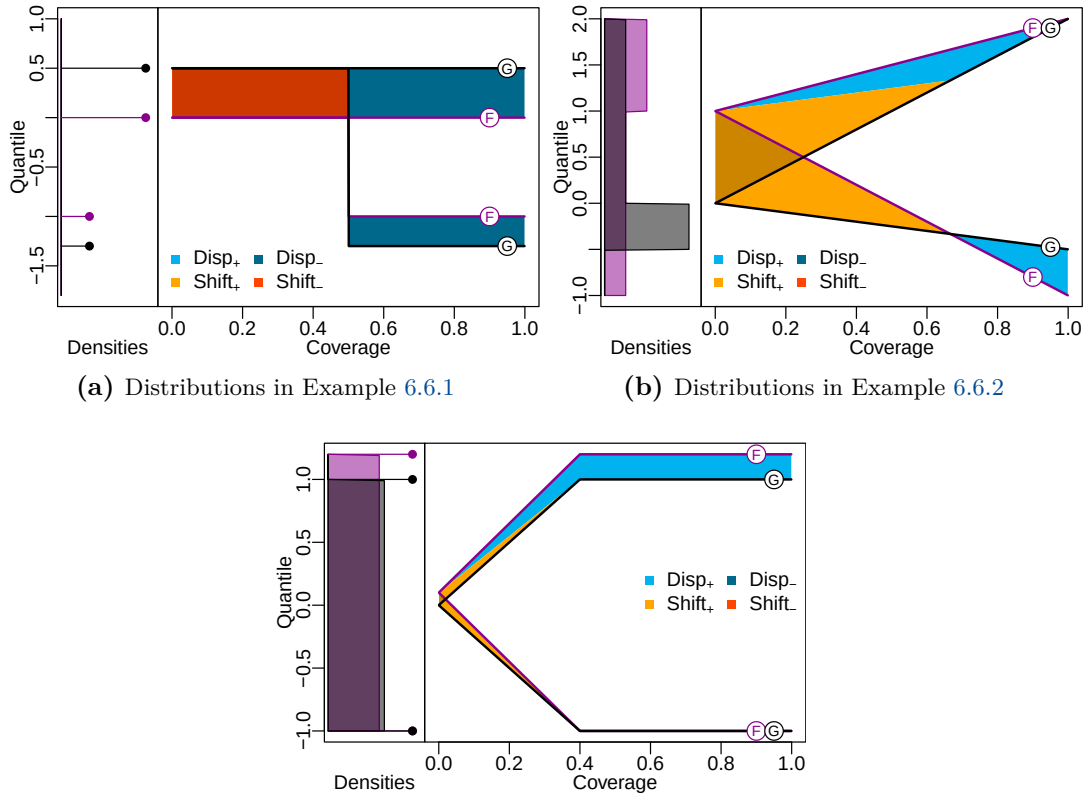
We omit the tedious derivations of the closed-form expressions for normal distributions given here and in the main text.

D.4. Counterexamples

This section contains various counterexamples that illustrate that certain restrictions in our theoretical results from Sections 6.3–6.4 are required.

D.5. Counterexamples of decomposition comparisons

We start to illustrate the necessity of the *symmetry* condition in Theorem 6.3.10. The following counterexample shows that the corresponding inequality (6.23) is not guaranteed to hold for asymmetric distributions, even if we focus attention on location-scale families as formally defined in Proposition 6.3.4.



(a) Distributions in Example 6.6.1

(b) Distributions in Example 6.6.2

(c) Distributions in Example 6.6.3

Figure D.2.: Illustrations of the AVM decompositions for distributions F and G given in Examples 6.6.1–6.6.3. See Figure 6.1 for a detailed description of the plots.

Example 6.6.1. Consider the distributions F and G that are illustrated in panel (a) of Figure D.2. The distribution F places $\frac{1}{4}$ probability mass at -1 and $\frac{3}{4}$ probability mass at 0 and G is given by the location-scale transformation $G(x) = F(\frac{x-0.5}{1.8})$. The first three p-th power p-Wasserstein distances between F and G and their decompositions are

$$\begin{aligned}\mathrm{WD}_1(F,G) &= \mathrm{Shift}_+^{\mathrm{WD}_1}(F,G) + \mathrm{Shift}_-^{\mathrm{WD}_1}(F,G) + \mathrm{Disp}_+^{\mathrm{WD}_1}(F,G) + \mathrm{Disp}_-^{\mathrm{WD}_1}(F,G)\\ &= 0 + 0.25 + 0 + 0.20,\\ \mathrm{WD}_2(F,G) &= 0 + 0.125 + 0 + 0.085, \qquad \mathrm{WD}_3(F,G) = 0 + 0.0625 + 0 + 0.038.\end{aligned}$$

Hence, the dispersion component accounts for about 44.4% of $\mathrm{WD}_1(F, G)$, which reduces to about 40.5% of $\mathrm{WD}_2(F, G)$ and about 37.8% of $\mathrm{WD}_3(F, G)$.

A corresponding example with continuous distributions can be obtained by slightly tilting the horizontal and vertical (jump) segments of the quantile functions in Figure D.2 (a). However, the given example is easy to grasp as the components are simply given by the areas of the three rectangles, where it is important to note that the red rectangle is counted twice for the shift component (compare to Figure 6.1). As the p-th power p-Wasserstein distance takes the p-th power of the height of the rectangles, the lower (smaller) rectangle contributing to the dispersion component is down-weighted in comparison to the other rectangles when increasing the power p, thereby reducing the relative weight of the dispersion component. Evidently, the relative weight of the dispersion component converges to $\frac{1}{3}$ from above as $p \to \infty$.

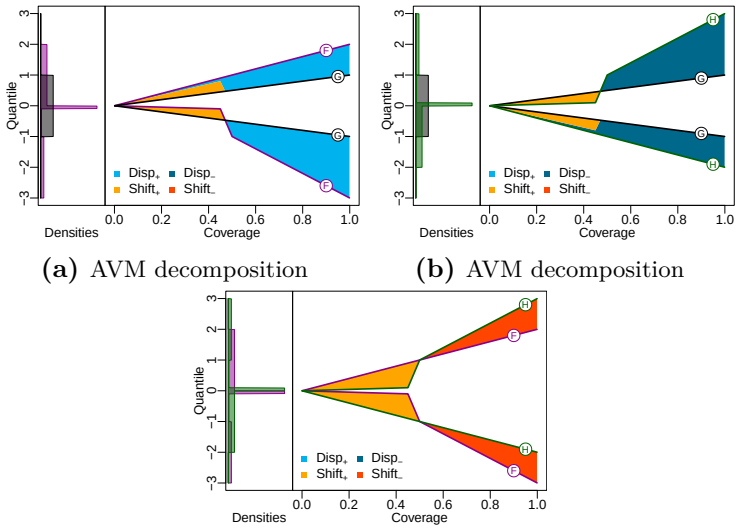
The following example illustrates that Theorem 6.3.10 and its inequality (6.23) do not hold for arbitrary (weakly) unimodal distributions, formally defined in Conjecture 6.3.12.

Example 6.6.2. Consider the distributions $F = 0.5 \times \mathcal{U}[-1, 1] + 0.5 \times \mathcal{U}[1, 2]$ and $G = 0.5 \times \mathcal{U}[-0.5, 0] + 0.5 \times \mathcal{U}[0, 2]$ illustrated in panel (b) of Figure D.2. The decompositions of the first three p-th power p-Wasserstein distances between F and G are

$$\begin{aligned}
\mathrm{WD}_1(F, G) &= \mathrm{Shift}_+^{\mathrm{WD}_1}(F, G) + \mathrm{Shift}_-^{\mathrm{WD}_1}(F, G) + \mathrm{Disp}_+^{\mathrm{WD}_1}(F, G) + \mathrm{Disp}_-^{\mathrm{WD}_1}(F, G) \\
&= 0.3333 + 0 + 0.125 + 0, \\
\mathrm{WD}_2(F, G) &= 0.2222 + 0 + 0.0694 + 0, \qquad \mathrm{WD}_3(F, G) = 0.1667 + 0 + 0.0469 + 0.
\end{aligned}$$

Hence, the dispersion component accounts for about 27.3% of $\mathrm{WD}_1(F, G)$, which reduces to about 23.8% of $\mathrm{WD}_2(F, G)$ and about 22.0% of $\mathrm{WD}_3(F, G)$. A counterexample with strongly unimodal distributions (that are strictly in- and decreasing left and right of the unique mode) can again be constructed by slightly tilting the respective horizontal lines in the density functions shown in Figure D.2 (b).

We turn now to generalizations of inequality (6.24) between relative (normalized) dispersion components of the Cramér distance and the area validation metric. Contrary to inequality (6.23), inequality (6.24) does not hold for arbitrary symmetric distributions, as illustrated by the following example.



(a) AVM decomposition **(b)** AVM decomposition

(c) AVM decomposition

Figure D.3.: Illustrations of AVM decompositions for the distributions given in Example 6.6.4. Here, the same nonzero components arise in the CD decomposition. See Figure 6.1 for a detailed description of the plots.

Example 6.6.3. Consider the distributions $F = \frac{3}{10} \times \delta_{-1} + \frac{1}{5} \times \mathcal{U}[-1, 0.1] + \frac{1}{5} \times \mathcal{U}[0.1, 1.2] + \frac{3}{10} \times \delta_{1.2}$ and $G = \frac{3}{10} \times \delta_{-1} + \frac{1}{5} \times \mathcal{U}[-1, 0] + \frac{1}{5} \times \mathcal{U}[0, 1] + \frac{3}{10} \times \delta_1$ illustrated in panel (c) of Figure D.2, where δ_y denotes the Dirac measure in $y \in \mathbb{R}$. The decompositions of the Cramér distance and the area validation metric between F and G are given by

$$\mathrm{CD}(F, G) = 0.002 + 0 + 0.019 + 0, \qquad \mathrm{AVM}(F, G) = 0.02 + 0 + 0.08 + 0.$$

Here, the dispersion accounts for only about 9% of the Cramér distance, whereas it accounts for 20% of the area validation metric.

D.6. Counterexamples for the order relations

The following example illustrates the necessity of the common support assumption in Proposition 6.4.6 by showing that the weak stochastic order is not guaranteed to be a transitive relation in general.

Example 6.6.4. Consider the three distributions

$$\begin{aligned} F &= \tfrac{1}{4} \times \mathcal{U}[-3,-1] + \tfrac{1}{40} \times \mathcal{U}[-1,-0.1] + \tfrac{9}{40} \times \mathcal{U}[-0.1,0] + \tfrac{1}{2} \times \mathcal{U}[0,3], \\ G &= \mathcal{U}[-2,2], \\ H &= \tfrac{1}{2} \times \mathcal{U}[-3,0] + \tfrac{9}{40} \times \mathcal{U}[0,0.1] + \tfrac{1}{40} \times \mathcal{U}[0.1,1] + \tfrac{1}{2} \times \mathcal{U}[1,2], \end{aligned}$$

illustrated in Figure D.3, which do not have a common support, but continuous quantile functions. Their respective Cramér distance decompositions are approximately given by

$$\begin{aligned} \mathrm{CD}(F,G) &\approx 0.1336 + 0 + 0.8664 + 0, \\ \mathrm{CD}(G,H) &\approx 0.1345 + 0 + 0 + 0.8655, \\ \mathrm{CD}(F,H) &\approx 0.5615 + 0.4385 + 0 + 0. \end{aligned}$$

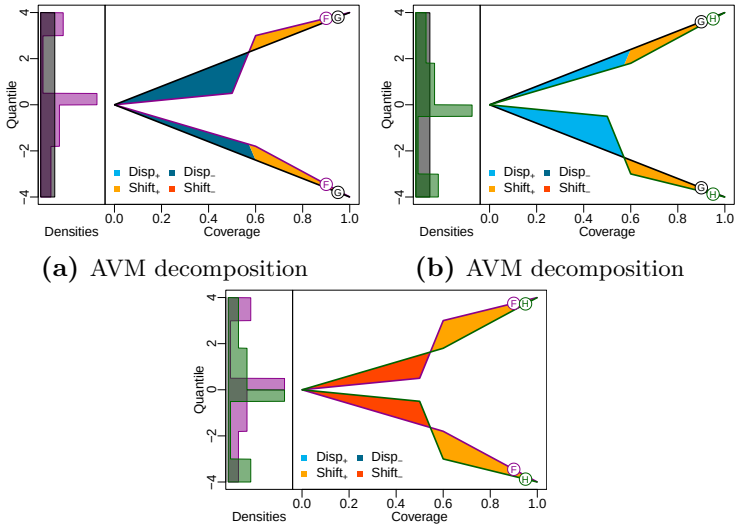
Thus, invoking part (a) of Theorem 6.4.4 for all three comparisons, we have $F \geq_{\mathrm{wS}} G \geq_{\mathrm{wS}} H$, but $F \not\geq_{\mathrm{wS}} H$. Hence, the weak stochastic order is not a transitive relation on arbitrary sets of distributions.

The following example shows that in contrast to the *weak* stochastic order (see Theorem 6.4.6), a common support is not sufficient to establish transitivity for the *relaxed* stochastic order.

Example 6.6.5. Consider the three distributions

$$\begin{aligned} F &= \tfrac{1}{5} \times \mathcal{U}[-4,-1.8] + \tfrac{3}{10} \times \mathcal{U}[-1.8,0] + \tfrac{1}{4} \times \mathcal{U}[0,0.5] + \tfrac{1}{20} \times \mathcal{U}[0.5,3] + \tfrac{1}{5} \times \mathcal{U}[3,4], \\ G &= \mathcal{U}[-4,4], \\ H &= \tfrac{1}{5} \times \mathcal{U}[-4,-3] + \tfrac{1}{20} \times \mathcal{U}[-3,-0.5] + \tfrac{1}{4} \times \mathcal{U}[-0.5,0] + \tfrac{3}{10} \times \mathcal{U}[0,1.8] + \tfrac{1}{5} \times \mathcal{U}[1.8,4], \end{aligned}$$

that are illustrated in Figure D.4, which have common support and continuous quantile functions. The quantile spread plots show unique nonzero shift components of the AVM, when comparing F and G (left), and G and H (middle), which implies a relaxed stochastic ordering, $F \geq_{\mathrm{rS}} G \geq_{\mathrm{rS}} H$ by Theorem 6.4.7. However, a comparison of F and H (right) yields two nonzero shift components, hence $G \not\geq_{\mathrm{rS}} H$ in relaxed stochastic order. Thus, the relaxed stochastic order is not a transitive relation on arbitrary sets



(a) AVM decomposition **(b)** AVM decomposition

(c) AVM decomposition

Figure D.4.: Illustrations of AVM decompositions for the distributions given in Example 6.6.5. See Figure 6.1 for a detailed description of the plots.

of distributions with common support. Notably, the CD decomposition produces two nonzero shift components, when comparing any of the three pairs, such that neither of the three comparisons yields a *weak* stochastic ordering by Theorem 6.4.4.

D.7. Approximation

In many cases, such as in the application in Section 6.5.2, the CDFs F and G under consideration are only partially known at a finite number of values or quantiles (that can even differ between F and G). Here, we outline two approximation strategies for such scenarios.

D.8. Distributions in a quantile format

An important special case of the setting described above arises if for both distributions quantile values at a fixed (and typically identical) set of levels are known. This is a common way of storing predictive distributions in a parsimonious format (see e.g., Cramer et al. 2022a). As often there is an interest in comparing predictions from different models or predictions issued at different time points (Amaral et al., 2024), we provide

some additional details on how to approximate the Cramér distance and compute its decomposition in this case. While we omit details, the presented formulations have proven most practical in our numerical experiments.

If the values of the generalized quantile functions F^{-1} and G^{-1} are known at levels $0 \leq \alpha_1 < \cdots < \alpha_K \leq 1$ and $0 \leq \beta_1 < \beta_2 < \ldots < \beta_L \leq 1$, respectively, the Cramér distance can be approximated as

$$\mathrm{CD}(F,G) \approx 2\sum_{i=1}^{K}\sum_{j=1}^{L} \frac{1+\mathbb{1}\{\alpha_i \neq \beta_j\}}{2}\frac{\alpha_{i+1}-\alpha_{i-1}}{2}\frac{\beta_{j+1}-\beta_{j-1}}{2}\chi\left(\alpha_i,\beta_j\right)\left|F^{-1}(\alpha_i)-G^{-1}\left(\beta_j\right)\right|,$$

where $\alpha_0 = \beta_0 = 0$ and $\alpha_{K+1} = \beta_{L+1} = 1$. If the quantiles bound central intervals (i.e., $\alpha_i = 1 - \alpha_{K+1-i}$ and $\beta_j = 1 - \beta_{L+1-j}$ holds for $i = 1, \ldots, K$ and $j = 1, \ldots, L$), then the components can be approximated similarly as

$$\begin{aligned}\mathrm{Shift}_{+}^{\mathrm{CD}}(F,G) \approx \sum_{i=1}^{\lceil K/2\rceil}\sum_{j=1}^{\lceil L/2\rceil}(1+\mathbb{1}\{\alpha_i \neq 0.5, \beta_j \neq 0.5\})\frac{\alpha_{i+1}-\alpha_{i-1}}{2}\frac{\beta_{j+1}-\beta_{j-1}}{2}\\ \cdot\left(\left[\min\left\{F^{-1}\left(\alpha_{K+1-i}\right)-G^{-1}\left(\beta_{L+1-j}\right), F^{-1}\left(\alpha_i\right)-G^{-1}\left(\beta_j\right)\right\}\right]_{+}+\left[F^{-1}\left(\alpha_i\right)-G^{-1}\left(\beta_{L+1-j}\right)\right]_{+}\right)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Disp}_{+}^{\mathrm{CD}}(F,G) \approx \sum_{i=1}^{\lceil K/2\rceil}\sum_{j=1}^{\lceil L/2\rceil}(1+\mathbb{1}\{\alpha_i \neq 0.5, \beta_j \neq 0.5\})\frac{1+\mathbb{1}\{\alpha_i \neq \beta_j\}}{2}\frac{\alpha_{i+1}-\alpha_{i-1}}{2}\frac{\beta_{j+1}-\beta_{j-1}}{2}\\ \cdot\left[\left(F^{-1}\left(\alpha_{K+1-i}\right)-G^{-1}\left(\beta_{L+1-j}\right)\right)-\left(F^{-1}\left(\alpha_i\right)-G^{-1}\left(\beta_j\right)\right)\right]_{+}\end{aligned}$$

The minus counterparts are approximated using the above formulas via the familiar symmetries $\mathrm{Shift}_{-}^{\mathrm{CD}}(F,G) = \mathrm{Shift}_{+}^{\mathrm{CD}}(G,F)$ and $\mathrm{Disp}_{-}^{\mathrm{CD}}(F,G) = \mathrm{Disp}_{+}^{\mathrm{CD}}(G,F)$. Formulas without the correction factor of $\frac{1}{2}$ that is applied in the case of equal quantile levels $\alpha_i = \beta_j$ tend to overestimate the CD and the corresponding dispersion components. The supplementary replication code provides an implementation of the approximation formulas, which sometimes served as a good alternative to using numeric integration techniques for the double integrals in (6.15)–(6.16).

D.9. Binned probability distributions

In the application from Section 6.5.2, the assigned bin probabilities together with the bin boundaries (i.e., the values separating the bins) identify the distribution and quantile functions at these boundary values. However, no further information on the quantile functions can be inferred from the responses.

In Figure 6.13, we illustrate the points where the distributions are known by the solid points in the quantile spread plots. To compute the Cramér and Wasserstein distances and their associated decompositions in this case, the quantile functions need to be approximated. For this, we adopt a simple linear interpolation between the known quantiles of the distributions as illustrated in Figure 6.13 with the straight lines connecting the points. In terms of the underlying histogram forecasts, this corresponds to uniformly distributing the probability mass within each bin.

In the tails (i.e., the open bins on either side), the histogram survey methodology does not capture any information on the distributions apart from the total probability mass beyond the most extreme bin boundaries. Hence, we adopt a conservative approach in order to avoid an overestimation of the distances and decomposition terms in the tails (where inherently little information is given): In the upper tail, we use the larger of the two lower bounds of the open bins to limit the support of the distributions, thereby shrinking this bin to a point mass for one of the distributions (as shown in Figure 6.13 in the density plots, which results in a horizontal segment in the quantile spread plots at large coverages). Analogously, we use the smaller of the two upper bounds of the open bins capturing the lower tail probabilities to limit the support of the distributions from below, thereby shrinking the left-most bin to a point mass for one of the distributions.

D.10. Derivation of the decompositions

A simple case distinction yields the following lemma, which is used to derive the decompositions.

Lemma 6.6.6. *For real numbers* $A, B \in \mathbb{R}$ *and* $[A]_+ := \max(A, 0)$*, we have*

$$[A]_+ + [B]_+ = [A + B]_+ + \min\{A, -B\}]_+ + [\min\{-A, B\}]_+.$$

Proof of Lemma 6.6.6. The statement of Lemma 6.6.6 follows from a simple case distinction:

- If $A, B > 0$, the statement is immediate as $[A]_+ + [B]_+ = A + B$ and $[A+B]_+ + [\min\{A, -B\}]_+ + [\min\{-A, B\}]_+ = A + B + 0 + 0$.
- If $A, B \leq 0$, then $[A]_+ + [B]_+ = 0 + 0$ and $[A+B]_+ + [\min\{A, -B\}]_+ + [\min\{-A, B\}]_+ = 0 + 0 + 0$.
- If $A > 0$ and $B \leq 0$, we get that $[A]_+ + [B]_+ = A + 0 = A$. If additionally $A \geq -B$, we get $[A+B]_+ + [\min\{A, -B\}]_+ + [\min\{-A, B\}]_+ = (A+B) + (-B) + 0 = A$. Similarly, if $A < -B$, we get $[A+B]_+ + [\min\{A, -B\}]_+ + [\min\{-A, B\}]_+ = 0 + A + 0 = A$.
- The result for $A \leq 0$ and $B > 0$ follows from the previous case by interchanging A and B.

□

Proof of Proposition 6.2.1. The result follows directly from (the proof of) Proposition 6.2.3 for $p = 1$. □

Proof of Proposition 6.2.3. As in (6.11)–(6.12), we denote by $z^{[p]} = \operatorname{sgn}(z) \cdot |z|^p$ the *signed p-th power* of a number $z \in \mathbb{R}$. Let

$$A = A(\alpha) = \left(F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\alpha}{2}\right) \right)^{[p]},$$
$$B = B(\alpha) = \left(F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right) \right)^{[p]},$$

denote the signed powers of the difference between the lower and upper ends of the central intervals of F and G, respectively. Then

$$\begin{aligned}
\mathrm{WD}_p(F,G) &= \int_0^1 \left|F^{-1}(\tau) - G^{-1}(\tau)\right|^p \mathrm{d}\tau \\
&= \int_0^{0.5} \left|F^{-1}(\tau) - G^{-1}(\tau)\right|^p \mathrm{d}\tau + \int_{0.5}^1 \left|F^{-1}(\tau) - G^{-1}(\tau)\right|^p \mathrm{d}\tau \\
&= \frac{1}{2}\int_0^1 \left|F^{-1}(\tfrac{1-\alpha}{2}) - G^{-1}(\tfrac{1-\alpha}{2})\right|^p \mathrm{d}\alpha + \frac{1}{2}\int_0^1 \left|F^{-1}(\tfrac{1+\alpha}{2}) - G^{-1}(\tfrac{1+\alpha}{2})\right|^p \mathrm{d}\alpha \\
&= \frac{1}{2}\int_0^1 \left|A(\alpha)\right| + \left|B(\alpha)\right| \mathrm{d}\alpha \\
&= \frac{1}{2}\int_0^1 [A(\alpha)]_+ + [-A(\alpha)]_+ + [B(\alpha)]_+ + [-B(\alpha)]_+ \,\mathrm{d}\alpha \\
&= \frac{1}{2}\int_0^1 \left[B(\alpha) - A(\alpha)\right]_+ + \left[A(\alpha) - B(\alpha)\right]_+ + 2\left[\min\{A(\alpha), B(\alpha)\}\right]_+ \\
&\qquad + 2\left[\min\{-A(\alpha), -B(\alpha)\}\right]_+ \mathrm{d}\alpha,
\end{aligned}$$

where the third equality is obtained by a simple change of variables and the last equality is obtained by applying Lemma 6.6.6 to $[A(\alpha)]_+ + [-B(\alpha)]_+$ and $[-A(\alpha)]_+ + [B(\alpha)]_+$.

Thus, the components in (6.11)–(6.12) are obtained by assigning the summands in the above equation as follows:

$$\begin{aligned}
\mathrm{Shift}_+^{\mathrm{WD}_p}(F,G) &= \frac{1}{2}\int_0^1 2\left[\min\{A(\alpha), B(\alpha)\}\right]_+ \mathrm{d}\alpha \\
&= \int_0^1 \left[\min\left\{\left(F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\alpha}{2}\right)\right)^{[p]}, \left(F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right)\right)^{[p]}\right\}\right]_+ \mathrm{d}\alpha, \\
\mathrm{Shift}_-^{\mathrm{WD}_p}(F,G) &= \frac{1}{2}\int_0^1 2\left[\min\{-A(\alpha), -B(\alpha)\}\right]_+ \mathrm{d}\alpha = \mathrm{Shift}_+^{\mathrm{WD}_p}(G,F),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Disp}_+^{\mathrm{WD}_p}(F,G) &= \frac{1}{2}\int_0^1 \left[B(\alpha) - A(\alpha)\right]_+ \mathrm{d}\alpha \\
&= \frac{1}{2}\int_0^1 \left[\left(F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right)\right)^{[p]} - \left(F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\alpha}{2}\right)\right)^{[p]}\right]_+ \mathrm{d}\alpha, \\
\mathrm{Disp}_-^{\mathrm{WD}_p}(F,G) &= \frac{1}{2}\int_0^1 \left[A(\alpha) - B(\alpha)\right]_+ \mathrm{d}\alpha = \mathrm{Disp}_+^{\mathrm{WD}_p}(G,F).
\end{aligned}$$

□

Proof of Proposition 6.2.4. We show the desired equality in three steps, which we treat individually below:

$$\begin{aligned}
\int_{-\infty}^{\infty} (F(x) - G(x))^2 \,\mathrm{d}x &= \int_0^1 \int_{-\infty}^{\infty} (F(x) - \mathbb{1}\{G^{-1}(\xi) \leq x\})^2 && (6.30)\\
&\qquad - (G(x) - \mathbb{1}\{G^{-1}(\xi) \leq x\})^2 \,\mathrm{d}x \,\mathrm{d}\xi && (6.31)\\
&= 2 \int_0^1 \int_0^1 (\mathbb{1}\{G^{-1}(\xi) \leq F^{-1}(\tau)\} - \tau)(F^{-1}(\tau) - G^{-1}(\xi)) && (6.32)\\
&\qquad - (\mathbb{1}\{G^{-1}(\xi) \leq G^{-1}(\tau)\} - \tau)(G^{-1}(\tau) - G^{-1}(\xi)) \,\mathrm{d}\tau \,\mathrm{d}\xi\\
&= 2 \int_0^1 \int_0^1 \chi(\tau, \xi) |F^{-1}(\tau) - G^{-1}(\xi)| \,\mathrm{d}\xi \,\mathrm{d}\tau. && (6.33)
\end{aligned}$$

The first step (6.31) essentially rewrites the Cramér distance as the divergence function associated with the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007). By noting that $G(z) = \int_0^1 \mathbb{1}\{\xi \leq G(z)\} \,\mathrm{d}\xi = \int_0^1 \mathbb{1}\{G^{-1}(\xi) \leq z\} \,\mathrm{d}\xi$, we obtain

$$\begin{aligned}
&\int_0^1 \int_{-\infty}^{\infty} (F(x) - \mathbb{1}\{G^{-1}(\xi) \leq x\})^2 - (G(x) - \mathbb{1}\{G^{-1}(\xi) \leq x\})^2 \,\mathrm{d}x \,\mathrm{d}\xi\\
&= \int_{-\infty}^{\infty} \int_0^1 F(x)^2 - 2F(x)\mathbb{1}\{G^{-1}(\xi) \leq x\} + \mathbb{1}\{G^{-1}(\xi) \leq x\}^2 - G(x)^2 + 2G(x)\mathbb{1}\{G^{-1}(\xi) \leq x\}\\
&\qquad - \mathbb{1}\{G^{-1}(\xi) \leq x\}^2 \,\mathrm{d}\xi \,\mathrm{d}x\\
&= \int_{-\infty}^{\infty} F(x)^2 - 2F(x) \int_0^1 \mathbb{1}\{G^{-1}(\xi) \leq x\} \,\mathrm{d}\xi + 2G(x) \int_0^1 \mathbb{1}\{G^{-1}(\xi) \leq x\} \,\mathrm{d}\xi - G(x)^2 \,\mathrm{d}x\\
&= \int_{-\infty}^{\infty} F(x)^2 - 2F(x)G(x) + 2G(x)G(x) - G(x)^2 \,\mathrm{d}x\\
&= \int_{-\infty}^{\infty} (F(x) - G(x))^2 \,\mathrm{d}x,
\end{aligned}$$

i.e., the first equality in (6.31).

The second step (6.32) is essentially equivalent to rewriting the CRPS in terms of quantile scores (Gneiting and Ranjan, 2011) and proceeds as suggested by Jordan (2016, Eq. (6.4) and (6.5)): To derive the second equality, it suffices to rewrite the inner integral in (6.32). We write $y = G^{-1}(\xi)$ for $\xi \in (0, 1)$ for ease of exposition. Note that the integral in (6.32) is made up of two similar terms, where the second term is rewritten analogously

by replacing $F^{-1}(\tau)$ with $G^{-1}(\tau)$ in the following. With these remarks, we obtain the second equality from

$$
\begin{aligned}
&\int_0^1 2(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau)(F^{-1}(\tau) - y)\,\mathrm{d}\tau \\
&= \int_0^1 \int_y^{F^{-1}(\tau)} 2(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau)\,\mathrm{d}x\,\mathrm{d}\tau \\
&= \int_0^1 \int_{-\infty}^{\infty} 2(\mathbb{1}\{x < F^{-1}(\tau)\} - \mathbb{1}\{x < y\})(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau)\,\mathrm{d}x\,\mathrm{d}\tau \\
&= \int_{-\infty}^{\infty} \int_0^1 2(\mathbb{1}\{x < F^{-1}(\tau)\} - \mathbb{1}\{x < y\})(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau)\,\mathrm{d}\tau\,\mathrm{d}x \\
&= \int_{-\infty}^{y} \int_0^1 \underbrace{2(-\mathbb{1}\{x \ge F^{-1}(\tau)\})(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau)}_{\neq 0,\text{ only if } F^{-1}(\tau) \le x \le y \text{ for } x \in (-\infty, y]}\,\mathrm{d}\tau\,\mathrm{d}x \\
&\quad + \int_y^{\infty} \int_0^1 \underbrace{2(\mathbb{1}\{x < F^{-1}(\tau)\})(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau)}_{\neq 0,\text{ only if } y \le x < F^{-1}(\tau) \text{ for } x \in [y, \infty)}\,\mathrm{d}\tau\,\mathrm{d}x \\
&= \int_{-\infty}^{y} \int_0^1 2(-\mathbb{1}\{F(x) \ge \tau\})(0 - \tau)\,\mathrm{d}\tau\,\mathrm{d}x + \int_y^{\infty} \int_0^1 2(\mathbb{1}\{F(x) < \tau\})(1 - \tau)\,\mathrm{d}\tau\,\mathrm{d}x \\
&= \int_{-\infty}^{y} \int_0^{F(x)} 2\tau\,\mathrm{d}\tau\,\mathrm{d}x + \int_y^{\infty} \int_{F(x)}^1 2(1 - \tau)\,\mathrm{d}\tau\,\mathrm{d}x \\
&= \int_{-\infty}^{y} F(x)^2\,\mathrm{d}x + \int_y^{\infty} (1 - F(x))^2\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} (F(x) - \mathbb{1}\{y \le x\})^2\,\mathrm{d}x.
\end{aligned}
$$

The third step (6.33) further simplifies the quantile-based representation of the CD to derive the concise formula presented in the proposition:

$$
\begin{aligned}
&2\int_0^1\int_0^1 (\mathbb{1}\{G^{-1}(\xi)\le F^{-1}(\tau)\}-\tau)(F^{-1}(\tau)-G^{-1}(\xi))\\
&\qquad -(\mathbb{1}\{G^{-1}(\xi)\le G^{-1}(\tau)\}-\tau)(G^{-1}(\tau)-G^{-1}(\xi))\,\mathrm{d}\tau\,\mathrm{d}\xi\\
&=2\int_0^1\int_0^\tau (\underbrace{\mathbb{1}\{G^{-1}(\xi)\le F^{-1}(\tau)\}}_{=1-\chi(\tau,\xi),\text{ as }\xi\le\tau}-\tau)(F^{-1}(\tau)-G^{-1}(\xi))-(1-\tau)(G^{-1}(\tau)-G^{-1}(\xi))\,\mathrm{d}\xi\\
&\quad+\int_\tau^1 (\underbrace{\mathbb{1}\{G^{-1}(\xi)\le F^{-1}(\tau)\}}_{=\chi(\tau,\xi),\text{ as }\xi\ge\tau}-\tau)(F^{-1}(\tau)-G^{-1}(\xi))+\tau(G^{-1}(\tau)-G^{-1}(\xi))\,\mathrm{d}\xi\,\mathrm{d}\tau\\
&=2\int_0^1 \underbrace{\int_0^\tau (1-\tau)(F^{-1}(\tau)-G^{-1}(\tau))\,\mathrm{d}\xi}_{=\tau(1-\tau)(F^{-1}(\tau)-G^{-1}(\tau))}-\int_0^\tau \underbrace{\chi(\tau,\xi)(F^{-1}(\tau)-G^{-1}(\xi))}_{\le 0\text{ as }\xi\le\tau}\,\mathrm{d}\xi\\
&\quad+\underbrace{\int_\tau^1 (-\tau)(F^{-1}(\tau)-G^{-1}(\tau))\,\mathrm{d}\xi}_{=-(1-\tau)\tau(F^{-1}(\tau)-G^{-1}(\tau))}+\int_\tau^1 \underbrace{\chi(\tau,\xi)(F^{-1}(\tau)-G^{-1}(\xi))}_{\ge 0\text{ as }\xi\ge\tau}\,\mathrm{d}\xi\,\mathrm{d}\tau\\
&=2\int_0^1\int_0^1 \chi(\tau,\xi)|F^{-1}(\tau)-G^{-1}(\xi)|\,\mathrm{d}\xi\,\mathrm{d}\tau.
\end{aligned}
$$

As all integrands are non-negative, changing the order of integration throughout the proof is not an issue by Fubini-Tonelli. No assumptions on the distributions are needed as long as integrals are allowed to be infinite. In the literature on proper scoring rules (in particular, the CRPS), existence of first moments is typically assumed to obtain a meaningful scoring rule that is strictly proper. For divergences, comparisons of two distributions without first moments might in some cases still yield a finite distance. Therefore, we dispense with such an assumption and provide a fully general proof.

□

Proof of Proposition 6.2.5. We start to define the following four differences of central interval endpoints

$$
\begin{aligned}
A=A(\alpha,\beta)&=F^{-1}\left(\tfrac{1-\alpha}{2}\right)-G^{-1}\left(\tfrac{1-\beta}{2}\right), & B=B(\alpha,\beta)&=F^{-1}\left(\tfrac{1+\alpha}{2}\right)-G^{-1}\left(\tfrac{1+\beta}{2}\right),\\
C=C(\alpha,\beta)&=F^{-1}\left(\tfrac{1-\alpha}{2}\right)-G^{-1}\left(\tfrac{1+\beta}{2}\right), & D=D(\alpha,\beta)&=F^{-1}\left(\tfrac{1+\alpha}{2}\right)-G^{-1}\left(\tfrac{1-\beta}{2}\right).
\end{aligned}
$$

We further define the function

$$\begin{aligned} f(\tau,\xi) &:= \chi(\tau,\xi)\,|F^{-1}(\tau) - G^{-1}(\xi)| \\ &= \mathbb{1}\left\{\operatorname{sgn}(\tau-\xi) \neq \operatorname{sgn}\left(F^{-1}(\tau) - G^{-1}(\xi)\right)\right\} |F^{-1}(\tau) - G^{-1}(\xi)|. \end{aligned}$$

for all $\tau, \xi \in [0,1]$, which arises as the integrand in (6.14). By the following case distinctions, we note that

$$\begin{aligned} f\left(\tfrac{1-\alpha}{2}, \tfrac{1-\beta}{2}\right) &= \begin{cases} 0, & \text{if } \alpha \leq \beta \text{ and } A \geq 0 \\ -A, & \text{if } \alpha \leq \beta \text{ and } A \leq 0 \\ A, & \text{if } \alpha \geq \beta \text{ and } A \geq 0 \\ 0, & \text{if } \alpha \geq \beta \text{ and } A \leq 0 \end{cases} \\ &= \mathbb{1}\{\alpha \geq \beta\}[A]_+ + \mathbb{1}\{\alpha \leq \beta\}[-A]_+. \end{aligned}$$

Similar considerations yield that

$$\begin{aligned} f\left(\tfrac{1-\alpha}{2}, \tfrac{1+\beta}{2}\right) &= [C]_+, \quad f\left(\tfrac{1+\alpha}{2}, \tfrac{1-\beta}{2}\right) = [-D]_+, \\ f\left(\tfrac{1+\alpha}{2}, \tfrac{1+\beta}{2}\right) &= \mathbb{1}\{\alpha \geq \beta\}[-B]_+ + \mathbb{1}\{\alpha \leq \beta\}[B]_+. \end{aligned}$$

Then, by a transformation of variables in the third equality, plugging in the above expressions in the fourth equality and by applying Lemma 6.6.6 in the fifth equality

below, we get that

$$
\begin{aligned}
\mathrm{CD}(F,G) &= 2\int_0^1\int_0^1 f(\tau,\xi)\,\mathrm{d}\tau\,\mathrm{d}\xi \\
&= 2\left[\int_0^{0.5}\int_0^{0.5} f(\tau,\xi)\,\mathrm{d}\tau\,\mathrm{d}\xi + \int_0^{0.5}\int_{0.5}^1 f(\tau,\xi)\,\mathrm{d}\tau\,\mathrm{d}\xi + \int_{0.5}^1\int_0^{0.5} f(\tau,\xi)\,\mathrm{d}\tau\,\mathrm{d}\xi + \int_{0.5}^1\int_{0.5}^1 f(\tau,\xi)\,\mathrm{d}\tau\,\mathrm{d}\xi\right] \\
&= 2\left[\frac{1}{4}\int_0^1\int_0^1 f\big(\tfrac{1-\alpha}{2},\tfrac{1-\beta}{2}\big)\,\mathrm{d}\alpha\,\mathrm{d}\beta + \frac{1}{4}\int_0^1\int_0^1 f\big(\tfrac{1-\alpha}{2},\tfrac{1+\beta}{2}\big)\,\mathrm{d}\alpha\,\mathrm{d}\beta\right. \\
&\quad \left. + \frac{1}{4}\int_0^1\int_0^1 f\big(\tfrac{1+\alpha}{2},\tfrac{1-\beta}{2}\big)\,\mathrm{d}\alpha\,\mathrm{d}\beta + \frac{1}{4}\int_0^1\int_0^1 f\big(\tfrac{1+\alpha}{2},\tfrac{1+\beta}{2}\big)\,\mathrm{d}\alpha\,\mathrm{d}\beta\right] \\
&= \frac{1}{2}\int_0^1\int_0^1 \mathbb{1}\{\alpha\geq\beta\}[A]_+ + \mathbb{1}\{\alpha\leq\beta\}[-A]_+ + [C]_+ + [-D]_+ + \mathbb{1}\{\alpha\geq\beta\}[-B]_+ + \mathbb{1}\{\alpha\leq\beta\}[B]_+\,\mathrm{d}\alpha\,\mathrm{d}\beta \\
&= \frac{1}{2}\int_0^1\int_0^1 \mathbb{1}\{\alpha\geq\beta\}\Big([A]_+ + [-B]_+\Big) + \mathbb{1}\{\alpha\leq\beta\}\Big([-A]_+ + [B_+]\Big) + [C]_+ + [-D]_+\,\mathrm{d}\alpha\,\mathrm{d}\beta \\
&= \frac{1}{2}\int_0^1\int_0^1 \mathbb{1}\{\alpha\geq\beta\}\Big([A-B]_+ + [\min\{A,B\}]_+ + [\min\{-A,-B\}]_+\Big) \\
&\qquad + \mathbb{1}\{\alpha\leq\beta\}\Big([-A+B]_+ + [\min\{-A,-B\}]_+ + [\min\{A,B\}]_+\Big) + [C]_+ + [-D]_+\,\mathrm{d}\alpha\,\mathrm{d}\beta \\
&= \frac{1}{2}\int_0^1\int_0^1 \mathbb{1}\{\alpha\geq\beta\}[A-B]_+ + \mathbb{1}\{\alpha\leq\beta\}[B-A]_+ \\
&\qquad + [\min\{A,B\}]_+ + [\min\{-A,-B\}]_+ + [C]_+ + [-D]_+\,\mathrm{d}\alpha\,\mathrm{d}\beta.
\end{aligned}
$$

The components of the decomposition in (6.15)–(6.16) are then obtained by setting

$$
\begin{aligned}
\mathrm{Disp}_+^{\mathrm{CD}}(F,G) &= \frac{1}{2}\int_0^1\int_0^\beta [B-A]_+\,\mathrm{d}\alpha\,\mathrm{d}\beta \\
\mathrm{Shift}_+^{\mathrm{CD}}(F,G) &= \frac{1}{2}\int_0^1\int_0^1 [\min\{A,B\}]_+ + [C]_+\,\mathrm{d}\alpha\,\mathrm{d}\beta, \\
\mathrm{Disp}_-^{\mathrm{CD}}(F,G) &= \frac{1}{2}\int_0^1\int_0^\alpha [A-B]_+\,\mathrm{d}\beta\,\mathrm{d}\alpha, \\
\mathrm{Shift}_-^{\mathrm{CD}}(F,G) &= \frac{1}{2}\int_0^1\int_0^1 [\min\{-A,-B\}]_+ + [-D]_+\,\mathrm{d}\alpha\,\mathrm{d}\beta.
\end{aligned}
$$

Notice that for the two minus components with subscript '$-$', changing the roles of F and G merely changes the sign of A and B. □

D.11. Derivation of theoretical properties

We start by proving Propositions 6.3.7 and 6.3.8 as these help to simplify the proofs of some of the basic properties given in Section 6.3.1. Notice that the proofs of Propositions 6.3.7 and 6.3.8 do not require any of the results from Section 6.3.1 apart from the obvious symmetry given by Proposition 6.3.1.

D.12. Proofs of propositions on equivalence of nonzero components

Proof of Proposition 6.3.7. For the Wasserstein distances, the equivalence between nonzero dispersion components is clear, as the signed p-th power preserves nonzero values in the integrands. The following similar (but technical) argument for the plus component shows that the CD dispersion components are nonzero whenever the respective AVM dispersions are nonzero by the symmetry from Proposition 6.3.1.

Let $B = \{\alpha \mid \mathrm{Disp}^{\mathrm{AVM}}_{\alpha,+}(F,G) > 0\}$ be the set of all values of the integration variable such that the integrand given in (6.8) is nonzero. The set B has positive Lebesgue measure if $\mathrm{Disp}^{\mathrm{AVM}}_{+}(F,G) > 0$. On the other hand, for $\beta \in B$, the inner integral in the CD dispersion,

$$\int_0^\beta \left[\left(F^{-1}\left(\tfrac{1+\alpha}{2}\right) - F^{-1}\left(\tfrac{1-\alpha}{2}\right)\right) - \left(G^{-1}\left(\tfrac{1+\beta}{2}\right) - G^{-1}\left(\tfrac{1-\beta}{2}\right)\right)\right]_+ \mathrm{d}\alpha,$$

is zero only if $\alpha \mapsto F^{-1}\left(\frac{1+\alpha}{2}\right) - F^{-1}\left(\frac{1-\alpha}{2}\right)$ is discontinuous at $\beta \in B$ (as otherwise, we can find a value $\beta' < \beta$ such that the integrand is strictly positive for all $\alpha \in [\beta', \beta]$, which yields a nonzero inner integral). By left-continuity, the quantile function F^{-1} is discontinuous at countably many values at most. Hence, the inner integral is nonzero for almost all $\beta \in B$, and integration across $\beta \in B$ yields a nonzero dispersion component.

Conversely, if the inner integral is nonzero in the CD dispersion, the integrand in the AVM dispersion component is also nonzero (because the integrand in the displayed term is increasing in α). Hence, the reverse implication also holds. □

Proof of Proposition 6.3.8. Here, we proceed similarly as in the proof of Proposition 6.3.7. For the Wasserstein distances, the equivalence between nonzero shift components is clear, as the signed p-th power preserves nonzero values in the integrands. A similar

(but technical) argument for the plus component shows that the CD shift components are nonzero whenever the respective AVM shifts are nonzero by the symmetry from Proposition 6.3.1.

Let $B = \{\alpha \mid \mathrm{Shift}^{\mathrm{AVM}}_{\alpha,+}(F,G) > 0\}$ be the set of all values of the integration variable such that the integrand given in (6.7) is nonzero. The set B has positive Lebesgue measure if $\mathrm{Shift}^{\mathrm{AVM}}_{+}(F,G) > 0$. On the other hand, for $\beta \in B$, the inner integral in the CD shift,

$$\int_0^1 \left[\min\left\{F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\beta}{2}\right),\ F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\beta}{2}\right)\right\}\right]_+ + \left[F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\beta}{2}\right)\right]_+ \mathrm{d}\alpha.$$

is zero only if F^{-1} is discontinuous at $\frac{1+\beta}{2}$ or $\frac{1-\beta}{2}$, which is only true for at most countably many values β by left-continuity of the quantile function F^{-1}. As the inner integral is nonzero for almost all $\beta \in B$, integration across $\beta \in B$ yields a nonzero shift component.

In contrast to the proof of Proposition 6.3.7, the converse is not true, because the integrand is not monotonic as illustrated by Example 6.3.9. □

D.13. Proofs of basic properties

Proof of Proposition 6.3.2. Note that $F_s^{-1}(z) = F^{-1}(z) + s$, and hence the shift s cancels out in the dispersion terms of the AVM and CD. □

Proof of Proposition 6.3.3. (a) We start by showing the claim for the CD. Suppose $m_F - m_G \leq 0$. For almost all pairs $(\alpha,\beta) \in (0,1)^2$, either $F^{-1}(\frac{1-\alpha}{2}) - G^{-1}(\frac{1-\beta}{2}) \leq 0$ or $F^{-1}(\frac{1+\alpha}{2}) - G^{-1}(\frac{1+\beta}{2}) = 2(m_F - m_G) - (F^{-1}(\frac{1-\alpha}{2}) - G^{-1}(\frac{1-\beta}{2})) \leq 0$ by symmetry. Therefore, the minimum across these two terms is for almost all $(\alpha,\beta) \in (0,1)^2$ not positive. Furthermore, we have $F^{-1}(\frac{1-\alpha}{2}) \leq m_F \leq m_G \leq G^{-1}(\frac{1+\beta}{2})$. Hence, the shift component

$$\mathrm{Shift}^{\mathrm{CD}}_{+}(F,G) = \frac{1}{2}\int_0^1\int_0^1 \Bigg[\underbrace{\min\left\{F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\beta}{2}\right), F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\beta}{2}\right)\right\}}_{\leq 0 \text{ for almost all } (\alpha,\beta)\in(0,1)^2}\Bigg]_+ + \Bigg[\underbrace{F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\beta}{2}\right)}_{\leq 0}\Bigg]_+ \mathrm{d}\alpha\,\mathrm{d}\beta = 0.$$

From Proposition 6.3.8 (shown above), it follows that $\mathrm{Shift}_+^{\mathrm{WD}_p}(F,G) = 0$ is zero as well for all $p \in \mathbb{N}$.

(b) By contraposition to (a), a positive shift ($\mathrm{Shift}_+^{\mathrm{D}}(F,G) > 0$) implies a corresponding ordering of medians ($m_F > m_G$).

To finish the proof, we show that a strict ordering of *unique* medians, $m_F > m_G$, implies a positive shift component of the area validation metric. (For the other distances, the shift component is then positive by Proposition 6.3.8.) By continuity of the quantile functions at $\frac{1}{2}$ (otherwise the medians would not be unique), there exists a neighborhood $\left[\frac{1}{2}-\delta, \frac{1}{2}+\delta\right] \subset (0,1)$ for some (small enough) $\delta \in (0,1)$ such that $F^{-1}(\tau) > G^{-1}(\tau)$ holds for all $\tau \in \left[\frac{1}{2}-\delta, \frac{1}{2}+\delta\right]$. Therefore, we obtain

$$\mathrm{Shift}_+^{\mathrm{AVM}}(F,G) \geq \int_0^{2\delta} \Big[\min\Big\{\underbrace{F^{-1}(\tfrac{1+\alpha}{2}) - G^{-1}(\tfrac{1+\alpha}{2})}_{>0,\text{ as } \frac{1+\alpha}{2} \leq \frac{1}{2}+\delta}, \underbrace{F^{-1}(\tfrac{1-\alpha}{2}) - G^{-1}(\tfrac{1-\alpha}{2})}_{>0,\text{ as } \frac{1-\alpha}{2} \geq \frac{1}{2}-\delta}\Big\}\Big]_+ \,\mathrm{d}\alpha > 0.$$

□

Proof of Proposition 6.3.4. By Propositions 6.3.7 and 6.3.8 (proved above), it suffices to show the equivalences in (a) and (b) for the AVM.

(a) The AVM dispersion component is given by

$$\begin{aligned}
\mathrm{Disp}_+^{\mathrm{AVM}}(F,G) &= \frac{1}{2}\int_0^1 \left[\left(F^{-1}\left(\tfrac{1+\alpha}{2}\right) - F^{-1}\left(\tfrac{1-\alpha}{2}\right)\right) - \left(G^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right)\right)\right]_+ \mathrm{d}\alpha \\
&= \frac{1}{2}\int_0^1 \left[(s_F - s_G)\left(H^{-1}\left(\tfrac{1+\alpha}{2}\right) - H^{-1}\left(\tfrac{1-\alpha}{2}\right)\right)\right]_+ \mathrm{d}\alpha \\
&= \frac{1}{2}[s_F - s_G]_+ \int_0^1 H^{-1}\left(\tfrac{1+\alpha}{2}\right) - H^{-1}\left(\tfrac{1-\alpha}{2}\right) \mathrm{d}\alpha,
\end{aligned}$$

where the latter integral is nonzero as H is non-degenerate, and hence the equivalence holds.

(b) Without loss of generality, we assume that the central median m_H of H is 0 (as we can always standardize an arbitrary reference distribution H by replacing it with $\overline{H}$ given by $\overline{H}^{-1} = H^{-1} - m_H$). Then, the location parameters ℓ_F and ℓ_G are the central medians, m_F and m_G, of F and G, respectively.

Now suppose that $m_F - m_G \leq 0$. If $s_F \geq s_G$, then $F^{-1}\left(\frac{1-\alpha}{2}\right) - G^{-1}\left(\frac{1-\alpha}{2}\right) = m_F - m_G + (s_F - s_G)H^{-1}\left(\frac{1-\alpha}{2}\right) \leq 0$ holds for all coverages $\alpha \in (0,1)$ as $0 = m_H \geq H^{-1}\left(\frac{1-\alpha}{2}\right)$. Otherwise, if $s_F \leq s_G$, then $F^{-1}\left(\frac{1+\alpha}{2}\right) - G^{-1}\left(\frac{1+\alpha}{2}\right) = m_F - m_G + (s_F - s_G)H^{-1}\left(\frac{1+\alpha}{2}\right) \leq 0$ holds for all $\alpha \in (0,1)$. Therefore, the minimum across these two terms is non-positive $\alpha \in (0,1)$. Hence, the shift component is zero:

$$\begin{aligned}\text{Shift}_{+}^{\text{AVM}}(F,G) &= \frac{1}{2}\int_0^1 \Big[\underbrace{\min\left\{F^{-1}\left(\tfrac{1+\alpha}{2}\right) - G^{-1}\left(\tfrac{1+\alpha}{2}\right), F^{-1}\left(\tfrac{1-\alpha}{2}\right) - G^{-1}\left(\tfrac{1-\alpha}{2}\right)\right\}}_{\leq 0}\Big]_+ \,\mathrm{d}\alpha \\ &= 0.\end{aligned}$$

Proposition 6.3.8 implies that $\text{Shift}_{+}^{\text{WD}_p}(F,G) = 0$ for all $p \in \mathbb{N}$.

The proof is finished by following the proof of Proposition 6.3.3 (b) word by word.

□

Proof of Theorem 6.3.6. (a) Let F and G be distributions from the same location-scale family with location parameters ℓ_F and ℓ_G, and scale parameters s_F and s_G, respectively. As in part (b) of Proposition 6.3.4, we assume w.l.o.g. that the location parameters match the central medians of the distributions, i.e., $\ell_F = m_F$ and $\ell_G = m_G$. Let $F_{m_G - m_F}$ be the shifted version of F that has location parameter l_G and scale s_F. By condition (6.21) and the symmetry in (6.17), we obtain

$$\text{Shift}_{\pm}^{\text{AVM}}(F_{m_G-m_F}, G) = 0,$$

and condition (6.20) (plus symmetry in (6.17)) yields

$$\text{AVM}(F_{m_G-m_F}, G) = \begin{cases} \text{Disp}_{+}^{\text{AVM}}(F_{m_G-m_F}, G), & \text{if } s_F \geq s_G, \\ \text{Disp}_{-}^{\text{AVM}}(F_{m_G-m_F}, G), & \text{if } s_F < s_G. \end{cases}$$

By invariance of dispersion components to shifts in (6.18), we obtain the unique dispersion terms

$$\begin{aligned}\mathrm{Disp}_{\pm}^{\mathrm{AVM}}(F,G) &= \mathrm{Disp}_{\pm}^{\mathrm{AVM}}(F_{m_G-m_F},G)\\ &= \begin{cases}\mathrm{AVM}(F_{m_G-m_F},G), & \text{if } \{\pm=+ \text{ and } s_F \geq s_G\} \text{ or } \{\pm=- \text{ and } s_F < s_G\},\\ 0, & \text{if } \{\pm=- \text{ and } s_F \geq s_G\} \text{ or } \{\pm=+ \text{ and } s_F < s_G\}.\end{cases}\end{aligned}$$

Invoking (6.21) (plus the symmetry in (6.17)) once more, we obtain the unique shift terms

$$\begin{aligned}&\mathrm{Shift}_{\pm}^{\mathrm{AVM}}(F,G)\\ &= \begin{cases}\mathrm{AVM}(F,G)-\mathrm{AVM}(F_{m_G-m_F},G), & \text{if } \{\pm=+ \text{ and } m_F \geq m_G\} \text{ or } \{\pm=- \text{ and } m_F < m_G\},\\ 0, & \text{if } \{\pm=- \text{ and } m_F \geq m_G\} \text{ or } \{\pm=+ \text{ and } m_F < m_G\}.\end{cases}\end{aligned}$$

The previous formulas for $\mathrm{Disp}_{\pm}^{\mathrm{AVM}}(F,G)$ and $\mathrm{Shift}_{\pm}^{\mathrm{AVM}}(F,G)$ provide a unique form for the decomposition terms after invoking the respective properties. As our decomposition given in Proposition 6.2.1 is one candidate for a decomposition that satisfies these properties, and there can only be one, we can conclude that the previously derived terms must equal our decomposition, which concludes this proof.

(b) The proof proceeds with analogous arguments as for the AVM in (a) with (6.21) replaced by (6.19).

□

D.14. Proofs of theorems on comparisons across distances

Proof of Theorem 6.3.10. Let $\widetilde{m} = m_F - m_G$ be the difference between the central medians of F and G (as defined prior to Proposition 6.3.3). We assume w.l.o.g. that $\widetilde{m} \geq 0$ (otherwise, we exchange F and G). Then, the minus shift components are zero by Proposition 6.3.3. Furthermore, we define $f(\tau) = F^{-1}(\tau) - G^{-1}(\tau)$ for $\tau \in (0,1)$. By symmetry of F and G, the function f is (almost surely) symmetric as well, that is, $f(\tau) = 2\widetilde{m} - f(1-\tau)$ holds for almost all $\tau \in (0,1)$. Let $A = \{\tau \mid f(1-\tau) > f(\tau) \geq 0\} \cup \{\tau < \frac{1}{2} \mid f(1-\tau) = f(\tau) \geq 0\}$. Then, the plus shift component of the p-Wasserstein distance is twice the integral of f over A as a change of variables yields $\mathrm{Shift}_{+}^{\mathrm{WD}_p}(F,G) = \int_0^1 \left[\min\left\{f(\tau)^{[p]}, f(1-\tau)^{[p]}\right\}\right]_+ \mathrm{d}\tau = 2\int_A |f(\tau)|^p \,\mathrm{d}\tau$. Let $\overline{A} = \{1-\tau \mid$

$\tau \in A\}$, $B = \{\tau \mid f(\tau) \geq 0 > f(1-\tau)\}$, and $\overline{B} = \{1-\tau \mid \tau \in B\}$. By the (almost sure) symmetry of f, the values τ such that $0 > f(\tau)$ and $0 > f(1-\tau) = 2\widetilde{m} - f(\tau)$ form a null set in $(0,1)$, and so does the complement of the disjoint union $A \cup \overline{A} \cup B \cup \overline{B}$ in $(0,1)$. Note that the symmetry of f yields the inequalities

$$\widetilde{m} \geq f(\tau) \quad \text{for almost all } \tau \in A \tag{6.34}$$

$$f(\tau) \geq 2\widetilde{m} \quad \text{for almost all } \tau \in B \tag{6.35}$$

$$f(1-\tau) \geq \widetilde{m} \quad \text{for almost all } \tau \in A \tag{6.36}$$

$$f(\tau) \geq -f(1-\tau) \quad \text{for almost all } \tau \in B \tag{6.37}$$

We use these four a.s. (almost sure with respect to the Lebesgue measure) inequalities to derive the inequality

$$\begin{aligned}
&\int_A \underbrace{|f(\tau)|}_{\leq \widetilde{m} \text{ a.s. by (6.34)}}^p \,\mathrm{d}\tau \cdot \left(\int_A |f(1-\tau)|^{p-1} \,\mathrm{d}\tau + \int_B |f(\tau)|^{p-1} + \underbrace{|f(1-\tau)|}_{\leq f(\tau) \text{ a.s. by (6.37)}}^{p-1} \,\mathrm{d}\tau \right) \\
&\leq \widetilde{m} \int_A |f(\tau)|^{p-1} \,\mathrm{d}\tau \cdot \frac{1}{\widetilde{m}} \left(\int_A \underbrace{\widetilde{m}}_{\leq f(1-\tau) \text{ a.s. by (6.36)}} |f(1-\tau)|^{p-1} \,\mathrm{d}\tau + \int_B \underbrace{2\widetilde{m}}_{\leq f(\tau) \text{ a.s. by (6.35)}} |f(\tau)|^{p-1} \,\mathrm{d}\tau \right) \\
&\leq \int_A |f(\tau)|^{p-1} \,\mathrm{d}\tau \cdot \left(\int_A |f(1-\tau)|^p \,\mathrm{d}\tau + \int_B |f(\tau)|^p \,\mathrm{d}\tau \right) \\
&\leq \int_A |f(\tau)|^{p-1} \,\mathrm{d}\tau \cdot \left(\int_A |f(1-\tau)|^p \,\mathrm{d}\tau + \int_B |f(\tau)|^p + |f(1-\tau)|^p \,\mathrm{d}\tau \right),
\end{aligned}$$

which is equivalent to the inequality

$$\begin{aligned}
&\int_A |f(\tau)|^p \,\mathrm{d}\tau \cdot \left(\int_{\overline{A}} |f(\tau)|^{p-1} \,\mathrm{d}\tau + \int_{B \cup \overline{B}} |f(\tau)|^{p-1} \,\mathrm{d}\tau \right) \\
&\leq \int_A |f(\tau)|^{p-1} \,\mathrm{d}\tau \cdot \left(\int_{\overline{A}} |f(\tau)|^p \,\mathrm{d}\tau + \int_{B \cup \overline{B}} |f(\tau)|^p \,\mathrm{d}\tau \right).
\end{aligned}$$

By adding the term $\int_A |f(\tau)|^p \,\mathrm{d}\tau \cdot \int_A |f(\tau)|^{p-1} \,\mathrm{d}\tau$ to both sides, we end up with the inequality

$$\int_A |f(\tau)|^p \,\mathrm{d}\tau \int_0^1 |f(\tau)|^{p-1} \,\mathrm{d}\tau \leq \int_A |f(\tau)|^{p-1} \,\mathrm{d}\tau \int_0^1 |f(\tau)|^p \,\mathrm{d}\tau,$$

which is equivalent to

$$\frac{\int_A |f(\tau)|^p \,\mathrm{d}\tau}{\int_0^1 |f(\tau)|^p \,\mathrm{d}\tau} \leq \frac{\int_A |f(\tau)|^{p-1} \,\mathrm{d}\tau}{\int_0^1 |f(\tau)|^{p-1} \,\mathrm{d}\tau}$$

or

$$\frac{\mathrm{Shift}_+^{\mathrm{WD}_p}(F,G)}{\mathrm{WD}_p(F,G)} \leq \frac{\mathrm{Shift}_+^{\mathrm{WD}_{p-1}}(F,G)}{\mathrm{WD}_{p-1}(F,G)}.$$

As the relative (normalized) shift component decreases as p increases, the sum of the relative (normalized) dispersion components increases with p, as formalized by inequality (6.23). □

Proof of Theorem 6.3.11. Let $\sigma_F \neq \sigma_G$ (otherwise all dispersion components are 0, and hence the inequalities are satisfied). With notation as used in the closed-form expressions given at the end of Section 6.3.1, inequality (6.24) is equivalent to

$$\begin{aligned} 0 &\leq \frac{\mathrm{CD}(F,G)}{\mathrm{Disp}_+^{\mathrm{CD}}(F,G) + \mathrm{Disp}_-^{\mathrm{CD}}(F,G)} - \frac{\mathrm{AVM}(F,G)}{\mathrm{Disp}_+^{\mathrm{AVM}}(F,G) + \mathrm{Disp}_-^{\mathrm{AVM}}(F,G)} \\ &= \frac{2\widetilde{\rho}\phi(\widetilde{\mu}/\widetilde{\rho}) + \widetilde{\mu}(2\Phi(\widetilde{\mu}/\widetilde{\rho}) - 1) - \sqrt{2}\phi(0)(\sigma_F + \sigma_G)}{2\widetilde{\rho}\phi(0) - \sqrt{2}\phi(0)(\sigma_F + \sigma_G)} - \frac{(2\Phi(\widetilde{\mu}/\widetilde{\sigma}) - 1)\widetilde{\mu} + 2\widetilde{\sigma}\phi(\widetilde{\mu}/\widetilde{\sigma})}{2\widetilde{\sigma}\phi(0)} =: f(\widetilde{\mu}). \end{aligned}$$

Note that $f(0) = 0$, and hence it suffices to show that the first derivative is non-negative, i.e., $f'(\widetilde{\mu}) \geq 0$ for $\widetilde{\mu} > 0$. The first derivative is given by (note that $\phi'(x) = -x\phi(x)$)

$$\begin{aligned} f'(\widetilde{\mu}) &= \frac{2\widetilde{\rho}\phi'(\widetilde{\mu}/\widetilde{\rho})/\widetilde{\rho} + \widetilde{\mu}(2\phi(\widetilde{\mu}/\widetilde{\rho})/\widetilde{\rho}) + (2\Phi(\widetilde{\mu}/\widetilde{\rho}) - 1)}{2\widetilde{\rho}\phi(0) - \sqrt{2}\phi(0)(\sigma_F + \sigma_G)} \\ &\quad - \frac{(2\Phi(\widetilde{\mu}/\widetilde{\sigma}) - 1) + (2\phi(\widetilde{\mu}/\widetilde{\sigma})/\widetilde{\sigma})\widetilde{\mu} + 2\widetilde{\sigma}\phi'(\widetilde{\mu}/\widetilde{\sigma})/\widetilde{\sigma}}{2\widetilde{\sigma}\phi(0)} \\ &= \frac{(2\Phi(\widetilde{\mu}/\widetilde{\rho}) - 1)}{2\widetilde{\rho}\phi(0) - \sqrt{2}\phi(0)(\sigma_F + \sigma_G)} - \frac{(2\Phi(\widetilde{\mu}/\widetilde{\sigma}) - 1)}{2\widetilde{\sigma}\phi(0)} \end{aligned}$$

Note that $f'(0) = 0$ and hence it suffices to show that the second derivative $f''(\widetilde{\mu})$ is non-negative for $\widetilde{\mu} > 0$. The second derivative is given by

$$\begin{aligned} f''(\widetilde{\mu}) &= \frac{2\phi(\widetilde{\mu}/\widetilde{\rho})/\widetilde{\rho}}{2\widetilde{\rho}\phi(0) - \sqrt{2}\phi(0)(\sigma_F + \sigma_G)} - \frac{2\phi(\widetilde{\mu}/\widetilde{\sigma})/\widetilde{\sigma}}{2\widetilde{\sigma}\phi(0)} \\ &= \frac{\phi(\widetilde{\mu}/\widetilde{\rho})\widetilde{\sigma}/\widetilde{\rho} - \phi(\widetilde{\mu}/\widetilde{\sigma})(\widetilde{\rho} - \frac{1}{\sqrt{2}}(\sigma_F + \sigma_G))/\widetilde{\sigma}}{\phi(0)\widetilde{\sigma}(\widetilde{\rho} - \frac{1}{\sqrt{2}}(\sigma_F + \sigma_G))} \end{aligned} \tag{6.38}$$

W.l.o.g. let $\sigma_F < \sigma_G = \sigma_F + \widetilde{\sigma}$. Note that $\widetilde{\rho} - \frac{1}{\sqrt{2}}(\sigma_F + \sigma_G) = \sqrt{\sigma_F^2 + (\sigma_F + \widetilde{\sigma})^2} - \frac{1}{\sqrt{2}}(2\sigma_F + \widetilde{\sigma})$ is zero if $\widetilde{\sigma} = 0$ and has positive derivative

$$\frac{\partial}{\partial \widetilde{\sigma}}(\sqrt{\sigma_F^2 + (\sigma_F + \widetilde{\sigma})^2} - \frac{1}{\sqrt{2}}(2\sigma_F + \widetilde{\sigma})) = \frac{\sigma_F + \widetilde{\sigma}}{\sqrt{\sigma_F^2 + (\sigma_F + \widetilde{\sigma})^2}} - \frac{1}{\sqrt{2}} = \frac{1}{\sqrt{\sigma_F^2/(\sigma_F + \widetilde{\sigma})^2 + 1}} - \frac{1}{\sqrt{2}}$$

if $\widetilde{\sigma} > 0$. Hence, the denominator in expression (6.38) is positive. Therefore, it suffices to show that the numerator in expression (6.38) is also non-negative. As $\widetilde{\rho} = \sqrt{\sigma_F^2 + \sigma_G^2} = \sqrt{(\sigma_F - \sigma_G)^2 + 2\sigma_F\sigma_G} > \sqrt{(\sigma_F - \sigma_G)^2} = \widetilde{\sigma}$ and hence $\phi(\widetilde{\mu}/\widetilde{\rho}) > \phi(\widetilde{\mu}/\widetilde{\sigma})$, it suffices to show that $\widetilde{\sigma}/\widetilde{\rho} \geq (\widetilde{\rho} - \frac{1}{\sqrt{2}}(\sigma_F + \sigma_G))/\widetilde{\sigma}$ or, equivalently,

$$\begin{aligned}
1 &\geq \widetilde{\rho}(\widetilde{\rho} - \frac{1}{\sqrt{2}}(\sigma_F + \sigma_G))/\widetilde{\sigma}^2 \\
&= \sqrt{(\sigma_F/\widetilde{\sigma})^2 + (\sigma_F/\widetilde{\sigma} + 1)^2}\left(\sqrt{(\sigma_F/\widetilde{\sigma})^2 + (\sigma_F/\widetilde{\sigma} + 1)^2} - (\sigma_F/\widetilde{\sigma} + (\sigma_F/\widetilde{\sigma} + 1))/\sqrt{2}\right) \\
&=: g(\sigma_F/\widetilde{\sigma}).
\end{aligned}$$

We can rewrite g as

$$\begin{aligned}
g(x) &= \sqrt{x^2 + (x+1)^2}\left(\sqrt{x^2 + (x+1)^2} - (x + (x+1))/\sqrt{2}\right) \\
&= \sqrt{(2x^2 + 2x) + 1}\left(\sqrt{(2x^2 + 2x) + 1} - \sqrt{(2x^2 + 2x) + 1/2)}\right) =: h(2x^2 + 2x).
\end{aligned}$$

As $g(0) = h(0) = 1 - 1/\sqrt{2} < 1$ and h has negative derivative

$$\begin{aligned}
h'(y) &= \frac{1/2}{\sqrt{y+1}}\left(\sqrt{y+1} - \sqrt{y+1/2}\right) + \sqrt{y+1}\left(\frac{1/2}{\sqrt{y+1}} - \frac{1/2}{\sqrt{y+1/2}}\right) \\
&= 1 - \frac{1}{2}\left(\frac{\sqrt{y+1/2}}{\sqrt{y+1}} + \frac{\sqrt{y+1}}{\sqrt{y+1/2}}\right) = 1 - \sqrt{\frac{y^2 + \frac{3}{2}y + \frac{3}{4}}{y^2 + \frac{3}{2}y + \frac{1}{2}}} < 0
\end{aligned}$$

for $y \geq 0$ (note that the numerator of the fraction under the root in the last term is clearly larger than the denominator), $g(x) = h(2x^2 + 2x)$ is also decreasing in x and, hence, the inequality $1 > g(0) > g(\sigma_F/\widetilde{\sigma})$ holds, which finishes the proof of the theorem. □

D.15. Derivation of connections to stochastic order relations

D.16. Proofs of connections to dispersive orders

Proof of Theorem 6.4.1. (a) We prove the equivalence for the AVM. Results for the other distances are an immediate consequence as nonzero dispersion components coincide by Proposition 6.3.7.

Clearly, $F \geq_{\mathrm{wD}} G$ implies $\mathrm{Disp}_{-}^{\mathrm{AVM}}(F, G) = 0$ as the defining condition implies that the integrand (6.8) is zero for all coverages α in $(0, 1)$.

We prove the reverse implication by contraposition. If $F \ngeq_{\mathrm{wD}} G$, then there exists a $\tau \in (\frac{1}{2}, 1)$ such that $F^{-1}(\tau) - F^{-1}(1-\tau) < G^{-1}(\tau) - G^{-1}(1-\tau)$. By continuity of the quantile functions there is a neighborhood $B \subset (\frac{1}{2}, 1)$ of τ such that the inequality $F^{-1}(\xi) - F^{-1}(1-\xi) < G^{-1}(\xi) - G^{-1}(1-\xi)$ holds for all $\xi \in B$. Substitution with $\tau = \frac{1+\alpha}{2}$ in the integral for the minus dispersion component (as in (6.10) with F and G switched) and rearranging terms yields

$$\mathrm{Disp}_{-}^{\mathrm{AVM}}(F, G) \geq \int_B \left[\left(G^{-1}(\tau) - G^{-1}(1-\tau)\right) - \left(F^{-1}(\tau) - F^{-1}(1-\tau)\right)\right]_+ \mathrm{d}\tau > 0.$$

(b) The equivalence is an immediate consequence of (a) as the strict ordering $F >_{\mathrm{wD}} G$ is given by the two conditions $F \geq_{\mathrm{wD}} G$ and $F \nleq_{\mathrm{wD}} G$, which are equivalent to $\mathrm{Disp}_{-}^{\mathrm{D}}(F, G) = 0$ and $\mathrm{Disp}_{+}^{\mathrm{D}}(F, G) > 0$, respectively (where the latter invokes the symmetry from Proposition 6.3.1).

□

Proof of Proposition 6.4.2. (a) Clearly, a dispersive ordering implies (6.25) for all $\tau \in (0.5, 1)$.

(b) If $F >_{\mathrm{D}} G$, then there exist $0 < \xi < \tau < 1$ such that the defining inequality

$$F^{-1}(\tau) - F^{-1}(\xi) > G^{-1}(\tau) - G^{-1}(\xi)$$

is strict. Since the inequality

$$F^{-1}(1-\xi) - F^{-1}(1-\tau) \geq G^{-1}(1-\xi) - G^{-1}(1-\tau)$$

also holds (as $1-\tau < 1-\xi$), we get the strict inequality

$$\underbrace{F^{-1}(\tau) - F^{-1}(1-\tau)}_{=:f(\tau)} - \underbrace{(F^{-1}(\xi) - F^{-1}(1-\xi))}_{=:f(\xi)} > \underbrace{G^{-1}(\tau) - G^{-1}(1-\tau)}_{=:g(\tau)} - \underbrace{(G^{-1}(\xi) - G^{-1}(1-\xi))}_{=:g(\xi)}. \tag{6.39}$$

We now consider a case distinction:

- In the case of $\tau > \xi \geq \frac{1}{2}$, inequality (6.39) yields

$$f(\tau) - \underbrace{(f(\xi) - g(\xi))}_{\geq 0 \text{ by } F >_{\mathrm{D}} G} > g(\tau) \quad \Longrightarrow \quad f(\tau) > g(\tau).$$

- In the case of $\tau \geq \frac{1}{2} > \xi$, either $f(\tau) > g(\tau)$ or $f(\tau) = g(\tau)$ (by the dispersive ordering $F >_{\mathrm{D}} G$). In the latter case, inequality (6.39) yields

$$-f(\xi) > \underbrace{g(\tau) - f(\tau)}_{=0} - g(\xi) \quad \Longrightarrow \quad f(1-\xi) > g(1-\xi).$$

- Finally, in the case of $\frac{1}{2} > \tau > \xi$, inequality (6.39) yields

$$\underbrace{f(\tau) - g(\tau)}_{\leq 0 \text{ by } F >_{\mathrm{D}} G} - f(\xi) > -g(\xi) \quad \Longrightarrow \quad f(1-\xi) > g(1-\xi).$$

Hence, in each of the three cases, there exists a $\tau \in (0.5, 1)$ such that the inequality in (6.25) is strict, which results in a strict weak dispersive ordering $F >_{\mathrm{wD}} G$.

□

D.17. Proofs of connections to stochastic orders

Proof of Theorem 6.4.4. Arguments similar to those in the proof of Theorem 6.4.1 yield the desired result, as we show in the following:

(a) It is easy to see that $F \geq_{\mathrm{wS}} G$ implies $\mathrm{Shift}_{-}^{\mathrm{CD}}(F, G) = 0$ as the defining condition implies that the integrand in (6.15) is zero for all coverage levels α and β in $(0, 1)$.

We prove the reverse implication by contraposition. If $F \not\geq_{\mathrm{wS}} G$, then there exists a pair $(\tau, \xi) \in (\frac{1}{2}, 1)^2$ such that $\max\left\{F^{-1}(\tau) - G^{-1}(\xi),\, F^{-1}(1-\tau) - G^{-1}(1-\xi)\right\} < 0$.

By continuity of the quantile functions there is a neighborhood $B \subset (\frac{1}{2}, 1)^2$ of (τ, ξ) such that the above inequality holds for all pairs $(\tau, \xi) \in B$. Substitution with $\tau = \frac{1+\beta}{2}$ and $\xi = \frac{1+\alpha}{2}$ in the integral for the shift component of the CD (as in (6.15) with F and G switched) yields

$$\begin{aligned}\text{Shift}_{-}^{\text{CD}}(F, G) \geq 2 \int_B \Big[\underbrace{\min\left\{ G^{-1}(\xi) - F^{-1}(\tau),\ G^{-1}(1-\xi) - F^{-1}(1-\tau) \right\}}_{>0} \Big]_+ \\ + \left[G^{-1}(\xi) - F^{-1}(1-\tau) \right]_+ \mathrm{d}\xi\, \mathrm{d}\tau > 0.\end{aligned}$$

Notice that the term in the lower line is non-negative such that omitting it still yields a strictly positive shift component.

(b) In analogy to the proof of Theorem 6.4.1 (b), the equivalence is an immediate consequence of part (a).

□

Proof of Proposition 6.4.5. (a) It is easy to see that stochastic ordering $F \geq_{\text{S}} G$ implies weak stochastic ordering $F \geq_{\text{wS}} G$: If $\tau \geq \xi$, we obtain $F^{-1}(\tau) - G^{-1}(\xi) \geq F^{-1}(\tau) - G^{-1}(\tau) \geq 0$ by the monotonicity of G^{-1} and stochastic ordering. Otherwise, if $\tau < \xi$, we obtain $F^{-1}(1-\tau) - G^{-1}(1-\xi) \geq F^{-1}(1-\tau) - G^{-1}(1-\tau) \geq 0$ from the stochastic ordering constraint. Hence, condition (6.27) is satisfied.

(b) Clearly, strict stochastic ordering implies strict weak stochastic ordering as there exists a τ such that one of the inequalities in (a) is strict.

□

Proof of Proposition 6.4.6. Clearly, the weak stochastic order is reflexive. To show the transitivity, consider three distributions $F \geq_{\text{wS}} G \geq_{\text{wS}} H$, and arbitrary $0.5 < \tau, \xi < 1$, which are fixed throughout the proof. The distribution F is larger than H in weak

stochastic order if the following inequality is satisfied

$$\begin{aligned} 0 &\leq \max\{F^{-1}(\tau) - H^{-1}(\xi), F^{-1}(1-\tau) - H^{-1}(1-\xi)\} \\ &= \max\{\underbrace{F^{-1}(\tau) - G^{-1}(\gamma)}_{A(\gamma)} + \underbrace{G^{-1}(\gamma) - H^{-1}(\xi)}_{C(\gamma)}, \\ &\qquad \underbrace{F^{-1}(1-\tau) - G^{-1}(1-\gamma)}_{B(\gamma)} + \underbrace{G^{-1}(1-\gamma) - H^{-1}(1-\xi)}_{D(\gamma)}\}, \end{aligned}$$

where $0.5 < \gamma < 1$. For any γ, at least one of the terms $C(\gamma)$ or $D(\gamma)$ is non-negative by definition of the weak stochastic ordering $G \geq_{\text{wS}} H$. Therefore, it suffices to show that there exists a $\gamma_0 \in [0.5, 1]$ such that both $A(\gamma_0)$ and $B(\gamma_0)$ are non-negative to prove the above inequality. To this end, let $\gamma_0 = \sup\{\gamma \mid A(\gamma) \geq 0\}$. Note that $\gamma_0 \geq 0.5$ as $F^{-1}(\tau) \geq G^{-1}(0.5)$ by weak stochastic ordering (which, by setting $\xi = 0.5$ in its definition, implies that either $F^{-1}(\tau) \geq G^{-1}(0.5)$ or $F^{-1}(1-\tau) \geq G^{-1}(0.5)$ and $F^{-1}(\tau) \geq F^{-1}(1-\tau)$ holds as $\tau > 0.5$.)

We now distinguish two cases: First, in the case of $\gamma_0 = 1$, the equality in the centered equation above extends to $\gamma = 1$, and we obtain $A(\gamma_0) = A(1) = 0$ and $B(\gamma_0) \geq 0$ as $F^{-1}(1-\tau) \geq G^{-1}(1-\gamma_0) = G^{-1}(0)$ because of the common support. Second, if $\gamma_0 < 1$, assume $B(\gamma_0) < 0$. Then there exists $\varepsilon > 0$ such that $B(\gamma_0 + \varepsilon) < 0$ by continuity of the quantile functions, and $A(\gamma_0 + \varepsilon) < 0$ by the definition of γ_0. This however contradicts $F \geq_{\text{wS}} G$, such that we can conclude that $B(\gamma_0) \geq 0$. As $A(\gamma_0) = 0$ holds by continuity, this concludes the proof. □

Proof of Theorem 6.4.7. Arguments similar to those in the proofs of Theorems 6.4.1 and 6.4.4 yield the desired results. □

Proof of Proposition 6.4.8. (a) Clearly, condition (6.27) implies condition (6.28).

(b) If τ satisfies condition (6.29), then $\max\{G^{-1}(\tau) - F^{-1}(\tau), G^{-1}(1-\tau) - F^{-1}(1-\tau)\} < 0$. Hence, $F \not\leq_{\text{rS}} G$.

□

Proof of Proposition 6.4.9. In the case of two symmetric distributions, F and G, with continuous quantile functions, a combination of Theorem 6.4.7 and Proposition 6.3.3

yields that $F \geq_{\mathrm{rS}} G$ is equivalent to an ordering of the medians, $F^{-1}(\frac{1}{2}) \geq G^{-1}(\frac{1}{2})$, which clearly correspond to a reflexive and transitive relation. □

D.18. List of climate models

Table D.1.: List of CMIP5 models used in our meteorological application together with the corresponding institute(s).

Climate model	Institute(s)
ACCESS1-0	Commonwealth Scientific and Industrial Research Organisation (CSIRO) and Bureau of Meteorology, Australia
BCC-CSM1-1	Beijing Climate Center / China Meteorological Administration, China
BNU-ESM	College of Global Change and Earth System Science, Beijing Normal University, China
CanESM2	Canadian Centre for Climate Modelling and Analysis, Canada 2 National Center for Atmospheric Research (NCAR), USA
CCSM4	National Center for Atmospheric Research (NCAR), USA
CESM1-BGC	National Center for Atmospheric Research (NCAR), USA
CMCC-CM	Centro Euro-Mediterraneo per i Cambiamenti Climatici, Italy
CNRM-CM5	Centre National de Recherches Météorologiques / Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, France
CSIRO-Mk3-6-0	CSIRO in collaboration with the Queensland Climate Change Centre of Excellence, Australia
EC-EARTH	EC-Earth consortium, Sweden
GFDL-CM3	Geophysical Fluid Dynamics Laboratory, USA
GFDL-ESM2G	Geophysical Fluid Dynamics Laboratory, USA
GFDL-ESM2M	Geophysical Fluid Dynamics Laboratory, USA
HadCM3	Met Office Hadley Centre, UK
HadGEM2-CC	Met Office Hadley Centre, UK
HadGEM2-ES	Met Office Hadley Centre, UK
INMCM4	Institute for Numerical Mathematics, Russia
IPSL-CM5A-LR	Institut Pierre Simon Laplace, France
IPSL-CM5A-MR	Institut Pierre Simon Laplace, France
IPSL-CM5B-LR	Institut Pierre Simon Laplace, France
MIROC4h	Atmosphere and Ocean Research Institute (AORI), National Institute for Environmental Studies (NIES) and Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Japan
MIROC5	AORI, NIES and JAMSTEC, Japan
MIROC-ESM	JAMSTEC, AORI and NIES, Japan
MIROC-ESM-CHEM	JAMSTEC, AORI and NIES, Japan
MPI-ESM-LR	Max Planck Institute for Meteorology, Germany
MPI-ESM-MR	Max Planck Institute for Meteorology, Germany
MPI-ESM-P	Max Planck Institute for Meteorology, Germany
MRI-CGCM3	Meteorological Research Institute, Japan
NorESM1-M	Norwegian Climate Centre, Norway

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

7.1. Introduction

Predictive modeling of infectious diseases has received considerable attention in recent years. This was fueled by the public health crises of COVID-19 (e.g., Cramer et al. 2022b; Bracher et al. 2021b) and mpox (e.g., Bleichrodt et al. 2024), but general principles had been developed previously, most prominently for seasonal influenza (Reich et al., 2019a). Disease forecasting is a broad field and three main types of predictive modeling tasks can be distinguished (Reich et al. 2022, see Figure 7.1).

- **Nowcasting** refers to the statistical correction of recent data points which are yet incomplete and subject to delayed additions. Nowcasts refer to recent rather than upcoming infection dynamics but are predictive in that they anticipate data revisions and reveal current trends.
- **Short-term forecasting** refers to unconditional predictions about the future course of an epidemic. Such predictions are feasible only for fairly short time periods, with appropriate prediction horizons depending on the indicator to predict.
- **Scenario predictions** are used to make statements about possible longer-term developments but are conditional on explicit assumptions that may or may not correspond to the future conditions encountered in the real world. For instance, scenarios may elucidate possible epidemic trajectories under various intervention strategies.

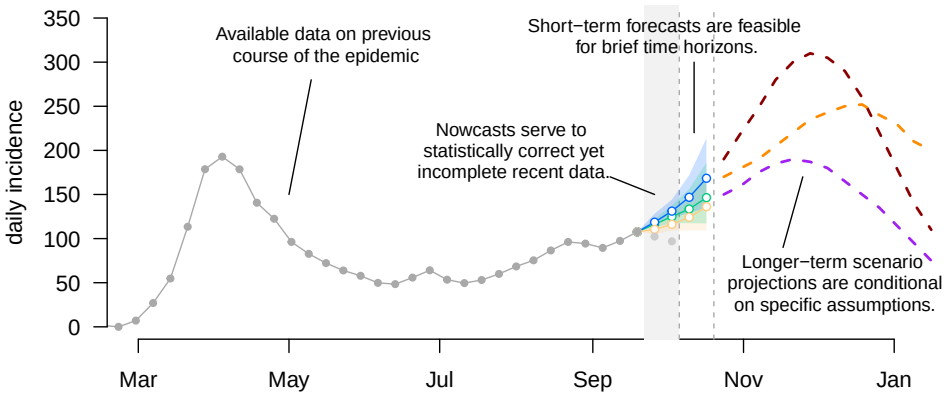


Figure 7.1.: Distinguishing nowcasting, short-term forecasting, and scenario modeling of infectious diseases.

Scenario modeling is conceptually different from the other two settings in that it yields statements about hypothetical settings rather than verifiable predictions. Nowcasting and short-term forecasting, on the other hand, ultimately boil down to the same task, namely probabilistic statements about disease incidence at various points in time. The availability of preliminary incidence values makes this more straightforward in nowcasting, but there is a natural transition between the two tasks. Somewhat surprisingly, nowcasting and short-term forecasting are rarely addressed jointly (an exception being Beesley et al. 2022).

For all three tasks, there is a growing consensus that multi-model approaches are particularly suitable. The presence of multiple distinct models enables more realistic assessments of the predictive uncertainty and can be the basis for ensemble forecasts, which have often been found to be more robust. Multi-model forecasting is often conducted in collaborative *Hubs* which collect contributions by participating teams. Recent efforts include e.g., the US COVID-19 Forecast and Scenario Modeling Hubs (Cramer et al., 2022b; Howerton et al., 2023), the European COVID-19 Forecast Hub (Sherratt et al., 2023) and the German COVID-19 Hospitalization Nowcast Hub (Wolffram et al., 2023). While the latter was explicitly limited to nowcasting, the forecasting platforms circumvented the nowcasting problem in various ways. As detailed e.g., in Bracher et al. (2021b), this can be done by aggregating incidence counts according to

the date of report rather than e.g., symptom onset. Alternatively, the most recent data points can be removed entirely to avoid dealing with delays, as done e.g., in a French hospitalization forecasting system (Paireau et al., 2022). Both approaches, however, blur or even discard valuable information on recent developments.

In this paper, we present a combined nowcasting and multi-model short-term forecasting system for hospitalizations due to severe acute respiratory infections (SARI) in Germany (November 2023–September 2024). Unlike forecast targets based on reporting dates, SARI hospitalizations are recorded by the date of hospital admission, making them susceptible to reporting delays and revisions, which necessitates special handling of the most recent, yet incomplete, data points. To this end, we develop a modular approach, where, rather than integrating a nowcasting step into each individual forecasting model, it is split off and handled by a separate statistical model. The resulting probabilistic nowcasts are fed back into a variety of forecasting methods. This includes a statistical time series model, a gradient boosting approach, and a neural network, which are moreover combined into an ensemble. The modular approach is practical as it avoids integrating nowcasting steps separately into conceptually diverse forecasting techniques. Moreover, it is helpful when historical data snapshots are not available for the entire period of interest; indeed, in our application, the first available data snapshot is from a release in September 2023, but the contained time series data reaches back to 2014. Splitting the training of nowcasting and forecasting models facilitates exploiting these disparate data sets in a straightforward manner. While the current paper is retrospective in nature, the developed approaches serve as a blueprint for a collaborative real-time nowcasting and forecasting system of infectious disease spread in Germany. As we will detail in Section 7.5, the so-called *RESPINOW Hub* was launched in fall 2024, and a prospective evaluation study on the 2024/25 season has recently been preregistered (Bracher and Wolffram, 2024).

A specific challenge in our application arises from the fact that the COVID-19 pandemic not only added to the general respiratory disease burden but also strongly impacted the dynamics of other respiratory diseases (see, e.g., Buchholz et al. 2023). This is true for the years 2020–2022 when the associated non-pharmaceutical interventions largely stopped the spread of other respiratory diseases, but also the following period, when

the immunity landscape was considerably different from earlier years. We will compare different approaches to using historical data from these periods for model fitting.

We find our forecasting models to be generally well-calibrated, though with rather wide uncertainty intervals surrounding peak weeks, and some difficulty in dealing with the double peak occurring in the test season. Similarly to previous studies, an ensemble approach achieves the best overall performance. Including a nowcasting step clearly improves forecasts relative to a procedure where the most recent data point is simply discarded. Indeed, the loss in forecasting performance relative to a hypothetical setting where the data are not subject to reporting delays turns out to be minor. This leads us to recommend the inclusion of nowcasting steps in infectious disease forecasting systems.

The remainder of the paper is structured as follows. Section 7.2 provides background information on SARI hospitalizations in Germany. In Section 7.3, we define the nowcasting and forecasting targets and present the methods employed for both tasks. Particular attention will be paid to the question of how to feed nowcast information into forecasting models while accounting for the arising uncertainties. In Section 7.4, we evaluate the resulting probabilistic forecasts visually and with a variety of metrics. In Section 7.5 we provide an outlook on the *RESPINOW Hub* before concluding with a discussion in Section 7.6.

7.2. The SARI hospitalization incidence

7.2.1. Definition and description

Respiratory disease activity in Germany is subject to a multitude of surveillance systems (see also Section 7.6). In the present paper, we are concerned with the hospitalization incidence for *severe acute respiratory infections* (SARI). Since fall 2014, data on such hospitalizations have been collected in the *ICOSARI* system operated by Robert Koch Institute (RKI; Buda et al. 2017; Tolksdorf et al. 2022a). They are publicly accessible via the RKI GitHub repository (`https://github.com/robert-koch-institut/SARI-Hospitalisierungsinzidenz`). The SARI hospitalization incidence is a *syndromic indicator*, i.e., the case definition is based on the symptoms patients present rather than laboratory testing for a specific pathogen. Specifically, a set of ICD-10 diagnostic codes

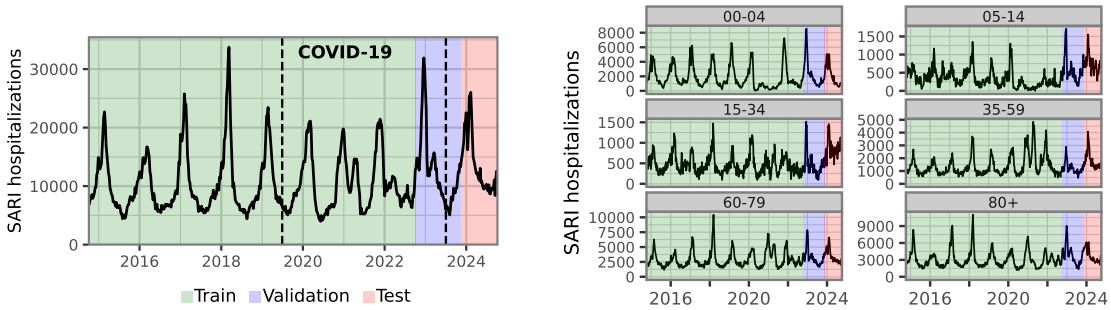


Figure 7.2.: Time series of weekly SARI hospitalizations in Germany, 2014–2024. Colors indicate the split of the data into training, validation, and test data; see details in Section 7.3.3. The portion labeled "COVID-19" is only included in the training set for part of our model specifications.

(J09–J22) is used, see Buda et al. (2017) for details. Data collection is carried out via a sentinel system that includes roughly 70 hospitals in 13 of the 16 German federal states. The system covers around 6% of all hospitalizations occurring in Germany. Based on an estimation of the catchment population covered by the sentinel sites, the SARI hospitalization incidence per 100,000 inhabitants can be estimated. Estimates in a weekly resolution (with weeks starting on Mondays) are made available both unstratified (00+) and for six age groups (0–4, 5–14, 15–34, 35–59, 60–79, 80+). In this paper, we rescale the estimated incidence to absolute count values.

The pooled and age group-wise incidence time series for the period 2014–2024 are displayed in Figure 7.2. Seasons we consider substantially affected by the acute phase of the COVID-19 pandemic are delimited by dashed vertical lines. Colors moreover indicate the split into training, validation, and test periods, see Section 7.3.3 for details. Especially in the age groups 05–14, 15–34, and 35–59, the test season displays rather unusual patterns, with consistently high incidences even in late spring and summer. In the very young and old age groups, this is less pronounced. As these age groups feature higher absolute numbers, the pooled incidence shown in the left panel likewise shows a more typical seasonal course.

We note that two of the considered forecasting models (`LightGBM` and `TSMixer`; Section 7.3.3) use an auxiliary data set on weekly outpatient consultations for acute respiratory infections (ARI). Details on this data set are provided in Appendix E.2 and a visualization is shown in Figure E.1.

7.2.2. Data revisions and reporting delays

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

As illustrated in the left panel of Figure 7.3, reporting delays lead to an artificial dip towards the end of the incidence time series data available in real time. Once data points have been completed over the following weeks, this dip disappears, and the actual trend becomes visible. For the SARI data, corrections become largely negligible after three weeks. The right panel shows the completeness of the data zero to four weeks after the initial release, by data release week. It can be seen that initial data releases on average contain roughly 75% of the hospitalizations (or, put differently, initial values are corrected upwards by roughly a third). Initial reporting completeness fluctuates somewhat over time. Between Christmas and New Year, no release occurs, meaning all hospitalizations from this period are reported with a delay.

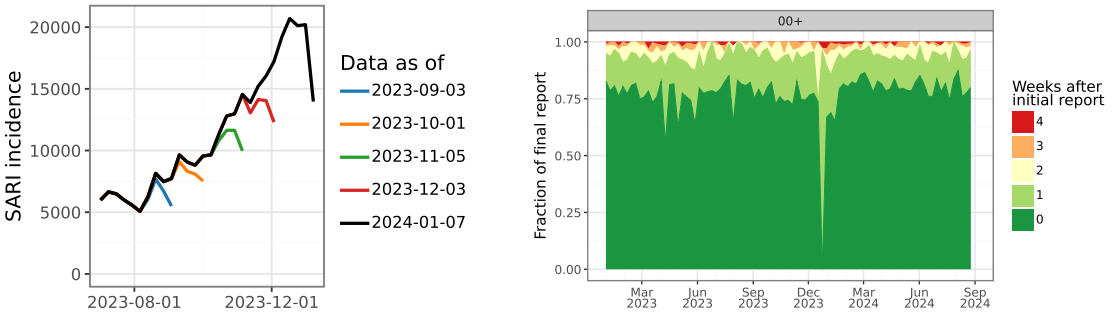


Figure 7.3.: Left: Illustration of data revisions in the SARI hospitalization incidence. Time series as available on different dates are shown in different colors, overlaid with more complete data in black. The apparent downward trend in the initial data versions is replaced by a continued upward trend in the revised data. Right: Completeness of SARI hospitalization data zero to four weeks after the first release, per week (2023–2024). In alignment with the nowcasting target definitions (see Section 7.3), we only consider delays up to four weeks.

7.3. Methods

7.3.1. Definition of nowcasting and forecasting tasks

Nowcasting addresses the statistical correction of partial/preliminary data. Real-time surveillance data are commonly completed or revised retrospectively. This is because by the time a new version of a surveillance data set is published, not all relevant reports will already have been received by the organization curating the data. Later versions of the data set will be updated with additional information. Nowcasting typically, but not necessarily, corresponds to upward correction of data to account for delayed reports. Forecasting concerns the future epidemiological development and thus time points for which not even partial data is currently available.

We generate nowcasts and forecasts in a weekly rhythm for the time period from 16 November 2023 through 12 September 2024, following the data release schedule on Thursdays. We skipped Thursday 28 December 2024 as no data release was available. This *test period* is highlighted in red in Figure 7.2. Counting from the day of data release (Thursday), the week ending on the preceding Sunday is indexed as *horizon* or *lead time* 0 weeks. Nowcasts, i.e., corrections of available preliminary data for e.g. reporting delays, are produced for weeks -3 through 0. Forecasts are generated for horizons 1 through 4. All

predictions are generated both for the total weekly number of SARI hospitalizations on the national level (aggregated across all ages) and stratified per age group. We note that the available SARI hospitalization incidence is actually only an estimate (see previous section). In practice, we neglect any uncertainty attached to these estimates and simply treat the estimates as the observable prediction target.

In the presence of data revisions, the definition of the prediction targets requires specific care. Based on experience from previous work (Wolffram et al., 2023), we define the final data version against which both nowcasts and forecasts are evaluated via a maximum reporting delay of $D = 4$ weeks. For each week, the respective data point used in the evaluation is thus set to the value available after four weeks of revisions (i.e., as published four weeks after the first data release containing a value for the respective week). This definition has the advantage of providing a well-defined target, with all observations in the evaluation period given the same amount of time for revisions. It is, however, unusual in that the time series used for evaluation is not identical to any specific public data release.

For each nowcast or forecast horizon, we collect predictive quantiles at levels 2.5%, 10%, 25%, 50%, 75%, 90%, and 97.5%. This storage format corresponds to that of various Forecast Hubs established during the COVID-19 pandemic (Cramer et al., 2022b; Wolffram et al., 2023). While it brings some constraints with it, we follow this convention as the present analysis aims to support the development of a new collaborative forecasting platform in Germany (see Section 7.5).

7.3.2. Nowcasting method and the coupling of nowcasting and forecasting

We separate the nowcasting and forecasting step and use a separate nowcasting model, which provides input for several forecasting models. While it may seem desirable to integrate nowcasting directly into each forecasting method, in practice this is hard to accommodate in many cases. Another challenge is the fact that data snapshots could only be partly recovered. We therefore split off the nowcasting from the forecasting task.

For nowcasting, we suggest a method that is based on the `simpleNowcast` first discussed in Wolffram et al. (2023, Supplementary Section E). It combines a straightforward

multiplication factor scheme with a parametric approach to estimate predictive uncertainty from past nowcast errors. Despite its simplicity, the approach showed performance comparable to more sophisticated approaches in Wolffram et al. (2023). In the present application, we need to adapt the original approach from daily to weekly data releases, which actually simplifies the technique as data release and now-/forecast schedules share the same frequency. The chosen simple format of the suggested nowcast technique has the key advantage that it allows to deal with very limited or missing information on strata of the full sample that characterize our data.

Point nowcast

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

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

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

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

is the number of hospitalizations still missing after d weeks.

In the following, we write X_t etc. for a random variable, x_t for the corresponding observation, and $\hat{x}_t$ for an estimated/imputed value. The hospitalizations per week and reporting delay as available at a given data release time t^* can be arranged into a *reporting triangle*, see Table 7.1.

Table 7.1.: Illustration of the reporting triangle for time t^* and $D = 4$. Quantities known at time t^* are shown in black, yet unknown quantities are shown in gray.

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

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

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

with a multiplication factor

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

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

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

and use it to impute

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

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

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

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

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

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

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

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

A slightly more formal explanation of how this relates to the estimation of a delay distribution from censored observations can be found in Wolffram et al. (2023). We note that this scheme would require some adaptations to deal with zeros in the reporting triangle, but these do not occur in our setting.

Nowcast uncertainty

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

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

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

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

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

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

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

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

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

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

We chose this very simple methodology because it is straightforward to adapt to a particularity of the nowcasting task at hand. In practice we encounter the problem that historical data snapshots are only available for the total SARI hospitalization incidence, but not the age-stratified time series (see Section 7.2.2). To nonetheless produce nowcasts at the stratified level we simplifyingly assume that the reporting delay distribution is identical across strata. The parameter estimates $\hat{\theta}_1, \ldots, \hat{\theta}_D$ are thus estimated from the pooled reporting triangles. The estimated overdispersion parameters $\hat{\psi}_0, \ldots, \hat{\psi}_{D-1}$ are likewise borrowed from the pooled fits.

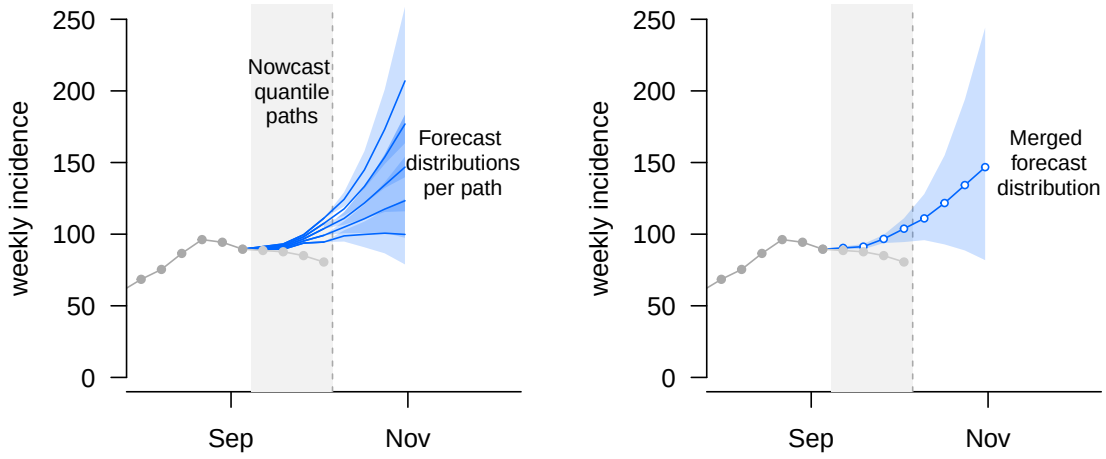


Figure 7.4.: Illustration of coupling between nowcasting and forecasting. A set of nowcast sample paths (blue lines in gray shaded area) is generated. Each of these is passed into a forecasting model to obtain predictive distributions for horizons 1 to 4. Results are then aggregated into overall forecast distributions via a linear pool (right panel).

Coupling of nowcasting and forecasting

For coupling the nowcast with the forecasting models, we propose the following model-agnostic approach to propagate nowcast uncertainty into forecasts (illustrated in Figure 7.4):

(a) Generate nowcast distributions for horizons -3 through 0 using a separate nowcasting model. For each horizon, quantiles at $N = 39$ different levels $0.025, 0.05, \ldots, 0.95, 0.975$ are generated.

(b) Translate these into 39 sample paths by assembling the predictive quantiles at identical levels for the four horizons.

(c) Feed each of these paths into the employed forecasting model to generate predictive distributions for horizons 1 through 4 (depending on the method these are samples or parametric distributions).

(d) Combine these predictive distributions by aggregating samples or averaging probability mass functions with linear pooling.

Step 2 is arbitrary in a sense as the distributions our nowcasting model returns for the various horizons are purely univariate and nothing is known about the dependence structure. However, in practice, the corrections at horizons -3 through -1 are minor and unlikely to have a major impact on predictions. It is therefore not crucial how exactly the nowcast paths are formed.

7.3.3. Forecasting methods

As SARI is not caused by an individual pathogen, it is not straightforward to model its dynamics mechanistically using classic compartmental (SIR-type) models. However, the SARI indicator is characterized by strong autocorrelations and, at least up to the COVID-19 pandemic, rather stable seasonal patterns. We therefore develop a suite of statistical models to exploit these regularities. As detailed in the following sections, some models can moreover exploit multivariate patterns across age groups (such as respiratory diseases often spreading from the younger to the older age groups) and information contained in auxiliary data streams.

Endemic-epidemic modeling: hhh4

The *endemic-epidemic* or `hhh4` model (after the associated function in the R package `surveillance`, Meyer et al. 2017) is a statistical time series model tailored to infectious disease surveillance data. While in principle it is capable of reflecting dependence structures across space or age groups, in our setting a simple univariate formulation for each stratum proved most robust. Denoting the incidence value (as absolute count value) in week t by X_t, the model is then defined as

$$X_t \mid \text{past} \sim \text{NegBin}(\text{mean} = \lambda_t, \text{disp} = \psi)$$
$$\lambda_t = \nu_t + \phi_t \times \sum_{d=1}^{D} w_d X_{t-d}.$$

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

In the model for the pooled time series we used the standard formulation

$$\nu_t = \alpha^{(\nu)} + \beta^{(\nu)} \times \sin(2\pi t/52) + \gamma^{(\nu)} \times \cos(2\pi t/52)$$
$$\phi_t = \alpha^{(\phi)} + \beta^{(\phi)} \times \sin(2\pi t/52) + \gamma^{(\phi)} \times \cos(2\pi t/52).$$

For the age-stratified forecasts, we further simplified this model and removed the intercept term ν_t (i.e., set it to zero). Already during the training period, retrospective forecasts from models including the intercept did not adapt well to changed magnitudes of incidence compared to earlier seasons. Especially in the age groups 05–14 and 35–59, this led to forecasts that were poorly aligned with the preceding data points. By removing the intercept this could be mitigated to a large degree.

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

The model fits are updated each week based on all historical data available (or, in a sensitivity analysis, excluding seasons strongly affected by the COVID-19 pandemic). Note that this also includes the corrected data points generated in the nowcasting step (see Section 7.3.2). Unlike the methods described in the two following subsections, no validation set is required, meaning that the distinction between the green and blue sections in Figure 7.2 is not relevant here. No additional data inputs other than the SARI incidences are used.

Gradient boosting: LightGBM

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

competition (Makridakis et al. (2022)), the model was one of the top models for predicting retail sales across multiple products and stores.

For our analysis, the model was retrained each week based on the historical data available and implemented in a multivariate fashion, allowing simultaneous prediction of all targets (i.e., different age groups and the national level). Weekly ARI numbers (see Appendix E.2) were included as a covariate. In addition to the lagged values of these two time series, the calendar week and the month of the subsequent week were also incorporated as input features. The last few observations that would remain incomplete in a real-time setting were excluded from the training process. They were subsequently replaced by nowcast paths to compute the forecasts as described in Section 7.3.2. Concerning hyperparameter selection (executed with W&B by Biewald 2020), after an initial random search to identify promising regions, an exhaustive grid search was performed over the refined hyperparameter space described in Table E.1 in the Appendix. To reduce computational requirements, the model was trained once on the training dataset and evaluated across all dates in the validation period (highlighted in green and blue in Figure 7.2). Due to the non-deterministic nature of the training process, we conducted training using 10 different random seeds and averaged the forecasts from these models (i.e., the predictive quantiles at each level) to obtain more robust results.

Deep learning model: TSMixer

The `TSMixer` architecture (as introduced in Chen et al. (2023a)) is a fully connected neural network specifically designed for time series forecasting. It utilizes a sequential mixing layer strategy that enables the model to capture both temporal dependencies and cross-feature interactions. As illustrated in Figure 7.5, the mixing layers are applied sequentially: first across the time dimension to model temporal patterns and then across the feature dimension to capture relationships between different variables. This approach allows the model to learn complex and non-linear relationships within the time series data. Compared to state-of-the-art transformer-based models, `TSMixer` often exhibits a simpler architecture, making it more computationally efficient and easier to train. Despite its relative simplicity, `TSMixer` has demonstrated competitive performance on a wide range of time series forecasting benchmarks, suggesting that its sequential mixing

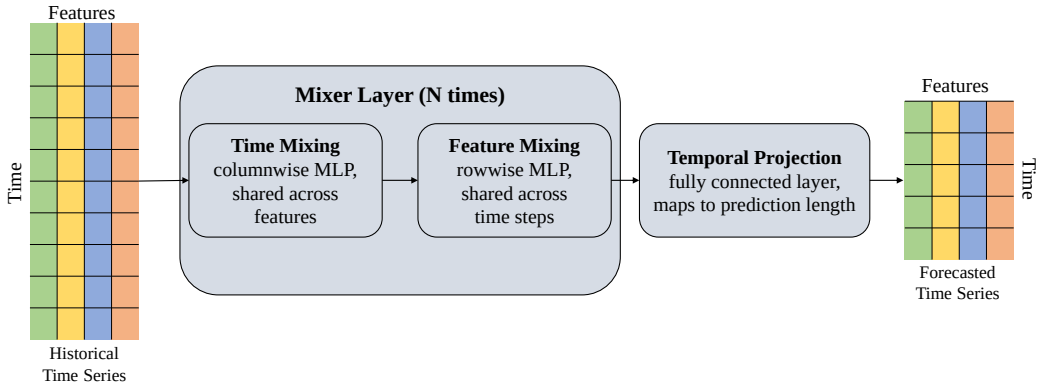


Figure 7.5.: Illustration of the `TSMixer` architecture, which is designed by stacking multi-layer perceptrons (MLPs). The mixing layers are applied repeatedly across the time dimension and the feature dimension to model both temporal patterns and interdependencies.

layer strategy is an effective approach for modeling temporal data. The model's ability to handle multivariate time series, as well as its potential for incorporating exogenous variables, makes it a versatile tool for various time series forecasting applications, such as infectious disease forecasting in our setting.

The implementation and training scheme follows that of `LightGBM` as described in the previous subsection.

7.3.4. The mean ensemble and reference models

For the `Ensemble`, the predictive quantiles were obtained as the arithmetic mean of the respective quantiles of the individual forecasts by the member models (`LightGBM`, `TSMixer`, and `hhh4`). This direct approach, also referred to as *Vincentization* (Genest, 1992) was chosen as other methods like the linear pool are not applicable when only a few predictive quantiles are available. As the present analyses shall serve as a blueprint

for a collaborative platform with quantile-based submissions (see Section 7.6), we work with this constraint and thus opt for the Vincentization approach.

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

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

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

7.3.5. Evaluation metrics

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

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

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

The WIS is negatively oriented, meaning that lower values are better. It can be seen as a probabilistic extension of the absolute error and approximates the commonly used continuous ranked probability score (CRPS, Gneiting et al. 2005). It is a proper scoring rule, thus incentivizing honest forecasting. To assess forecast calibration separately, the empirical coverage proportions of predictive 50% and 95% prediction intervals are reported.

7.4. Results

7.4.1. Visual inspection of nowcasts and forecasts

We start with a graphical assessment of nowcasts and forecasts. Figure 7.6 shows nowcasts and forecasts for the total SARI hospitalization incidence (pooled across age groups) issued by the `Ensemble` at nine different time points. To avoid overplotting, we use two separate panels and display the remaining time points in a set of figures in the Appendix (E.3). A detailed illustration of the nowcasts on the aggregate national level can also be found in Figure E.2. In Figure 7.6 at most instances, nowcasts (blue) are closely aligned with the completed data versions (black), but in some cases, discrepancies remain (e.g., for the second nowcast in the left panel). The nowcasting also successfully prevents forecasts from following spurious downward trends resulting from reporting delays. Forecasts are mostly well-aligned with the later observed trends, the exception being the first weeks of 2024 (see right panel). Here, the ensemble prediction implies that the peak has already occurred, failing to predict the second and higher peak. Such double peaks in close succession did not occur in any of the previous years, making this aspect hard to predict in a purely data-driven manner. The uncertainty intervals of nowcasts and forecasts are of adequate width to nonetheless cover the observed values in most instances. Especially around the peak, however, they become very wide, meaning that forecasts are less informative in these periods.

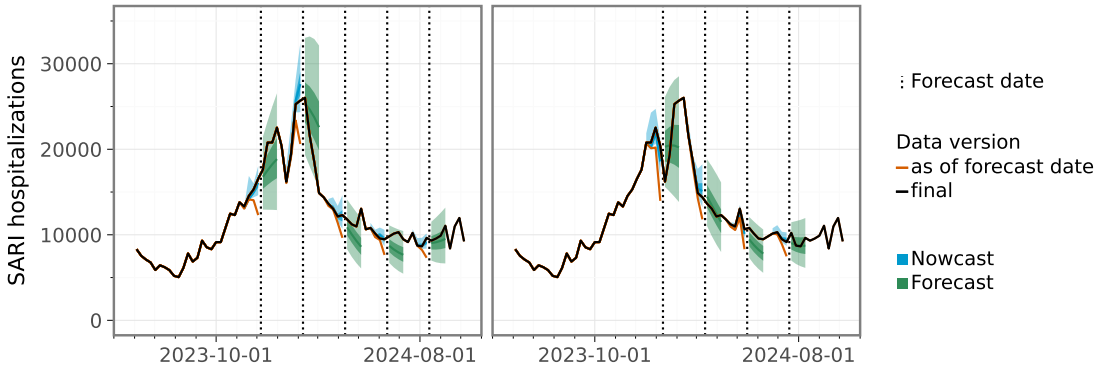


Figure 7.6.: Nowcasts and ensemble forecasts for the total SARI hospitalization incidence (pooled across age groups) at different forecast times. To avoid overplotting, we show the time series twice and overlay it with predictions issued at different times in the two panels. Figures covering all forecast dates are available in Figure E.3 in the Appendix.

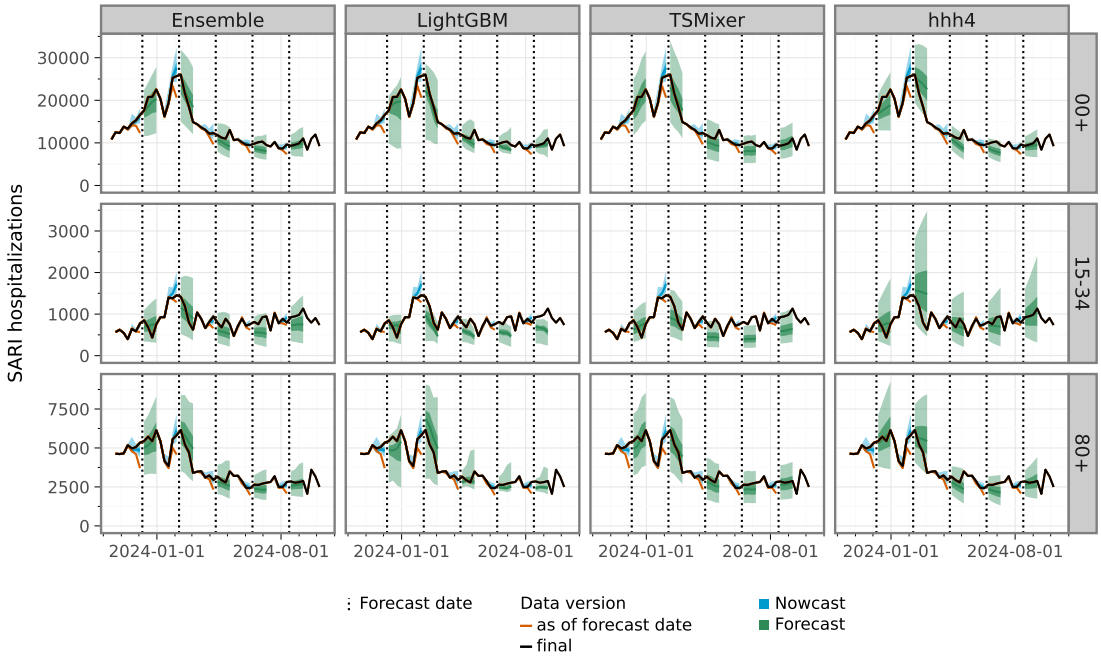


Figure 7.7.: Selected nowcasts and forecasts for the aggregate level 00+ and age groups 15-34 and 80+. To avoid overplotting, predictions for only five forecast times are shown. Figures covering all forecast dates are available for the `Ensemble` in Figure E.3 in the Appendix.

Selected predictions from individual models across age groups are displayed in Figure 7.7. As discussed in Section 7.2.1, age group 15–34, like some others, displayed unusual patterns in the 2023/24 season. Unlike in previous years, incidence stayed rather high throughout the spring and summer months. The `LightGBM` model struggles to adjust to this difference and keeps predicting a decline towards the usual levels (a similar pattern arises for `TSMixer`). The `hhh4` model with its simple autoregressive structure is better able to deal with this shift in magnitude. The difficulties of `LightGBM` and `TSMixer` are also inherited by the `Ensemble`. Similar patterns are also found for age group 05–14, and to a lesser degree for ages 35-59, while the remaining age groups have more typical seasonal courses. However, Figure 7.7 also illustrates some strengths of `LightGBM` and `TSMixer`, particularly at the national level (00+) and for older age groups (e.g., 80+). These models accurately capture the sharp decline following the second peak, whereas `hhh4` tends to produce more conservative forecasts.

7.4.2. Formal forecast evaluation

Aggregate-level nowcasts and forecasts

We complement the visual assessment with a more formal evaluation of forecast calibration and score-based performance. Figure 7.8 summarizes the national-level performance. Average WIS (across forecast dates) as well as the coverage fractions of 50% and 95% prediction intervals are displayed stratified by nowcast/forecast horizon. Little surprisingly, average scores increase with the horizon (i.e., performance decreases). For horizons 1 through 4, all models outperform the `Persistence` and `Historical` baseline models (with the exception of TSMixer at horizon 1). The `Ensemble` outperforms all individual models at all horizons, but the margin to `LightGBM` and `hhh4` is slim for short horizons. Interestingly, for horizon 4 this flips and the `TSMixer` model achieves performance close to the ensemble. The decomposition of the WIS indicates that `LightGBM` and `TSMixer` tend to underpredict, and that this tendency is inherited by the ensemble (this seems to be driven by the fact that the second peak was not anticipated, as well as the untypically high incidences of some age groups late in the season; see previous subsection). The `hhh4` and nowcasting models have more balanced components.

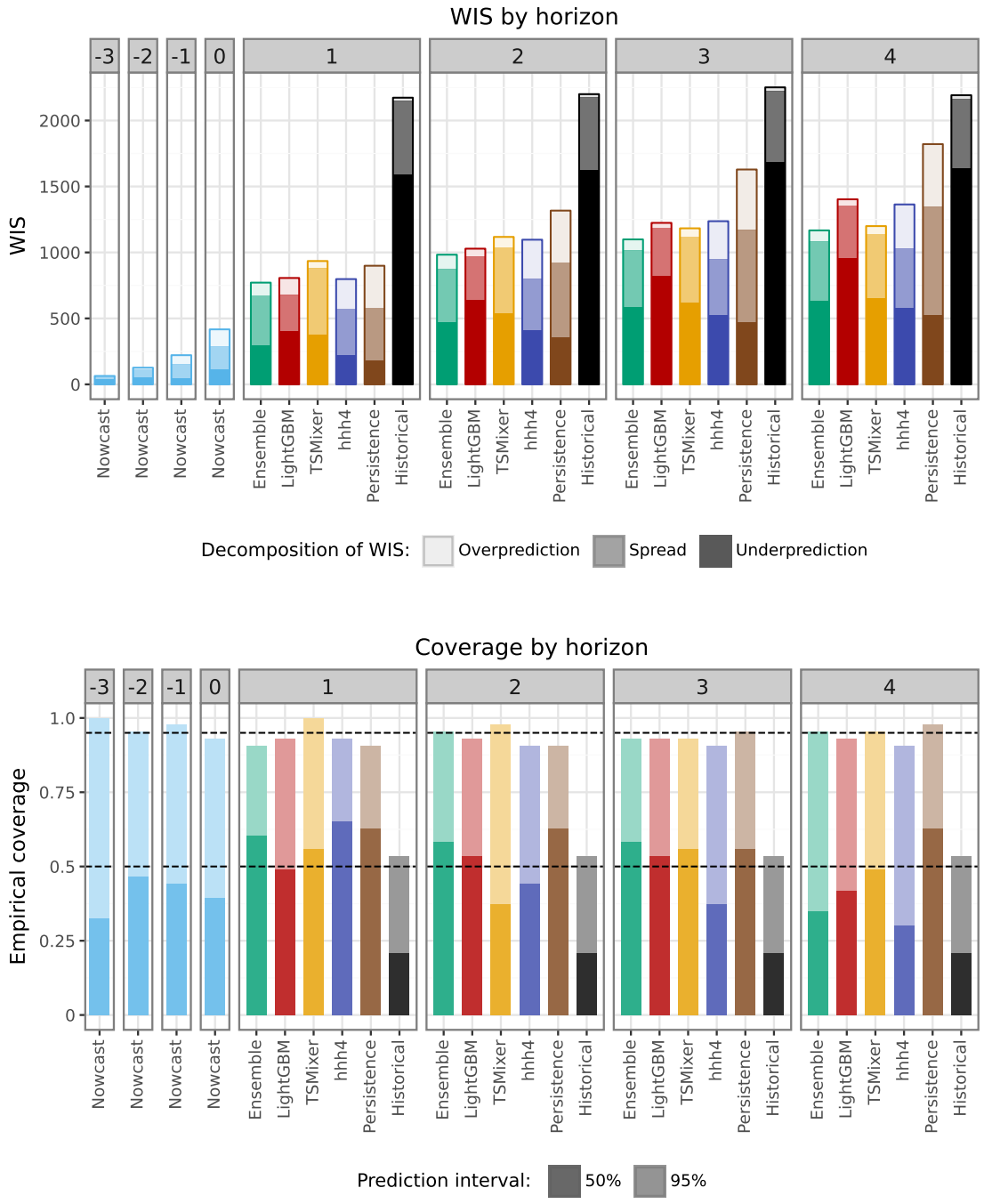


Figure 7.8.: Average WIS values (top) and empirical coverage rates (bottom) achieved by different models for the total SARI hospitalization incidence (pooled across age groups), stratified by nowcast/forecast horizon. The average scores are decomposed into components for overprediction, underprediction, and forecast spread.

A summary plot aggregating results across horizons is available in Figure E.4 (left panel) in the Appendix. While the `Ensemble` again has a little edge, the three member models `LightGBM`, `TSMixer` and `hhh4` are roughly on par.

Concerning the interval coverage rates (bottom panel in Figure 7.8), all models apart from the `Historical` baseline achieve close-to-nominal coverage.

Age-stratified nowcasts and forecasts

Figure 7.9 summarizes average results for age-stratified nowcasts and forecasts. The results in terms of average WIS are broadly consistent with those discussed in the previous section, with the ensemble again performing best across horizons and the individual models outperforming the baseline models in almost all instances. The `LightGBM` and `TSMixer` models again tend to underpredict, while the `hhh4` model features the most dispersed predictions.

The WIS stratified by age group (and aggregated by horizon), depicted in Figure 7.10, reveals that the aforementioned downward bias in `LightGBM` and `TSMixer` primarily originates from the age groups 05–14, 15–34, and 35–59. This can be attributed to the unusually high SARI incidence during the evaluation period (Figure 7.2), which did not follow the typical seasonal decline, as discussed previously. The score-based evaluation also confirms that the `hhh4` model performs particularly well in these age groups (in the 15–34 group even slightly outperforming the `Ensemble`). By contrast, the machine learning approaches had an edge in forecasting older age groups, potentially because they were able to leverage trends in younger age groups as leading indicators for older ones.

In terms of interval coverage (bottom panel of Figure 7.9), we observe that the nowcasts for horizons -1 and 0 are considerably overconfident. This is likely a consequence of the fact that only a few historical snapshots of age-stratified data were available, meaning that stratified nowcasts had to be based on aggregate-level snapshots (see Section 7.3.2).

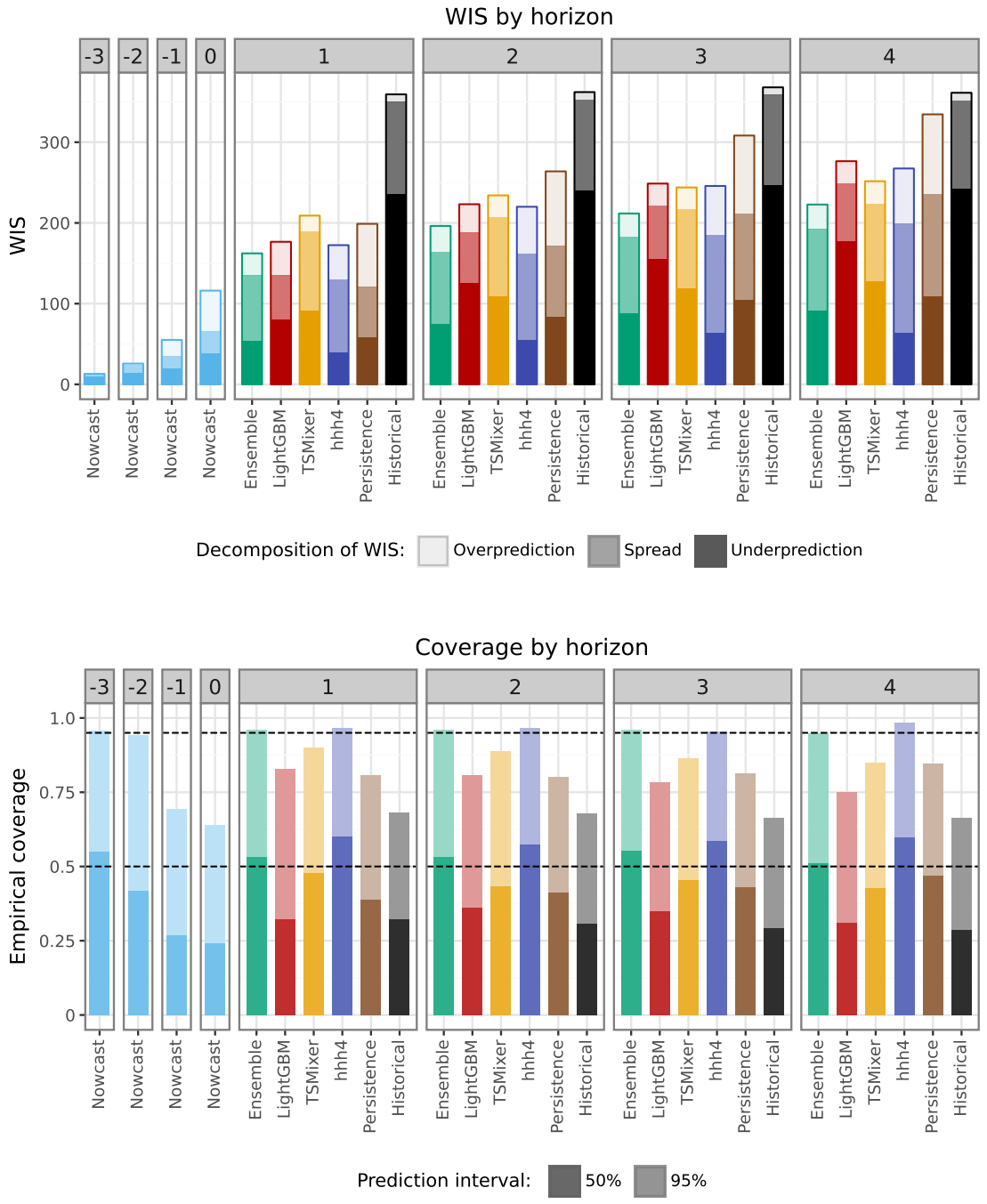


Figure 7.9.: Average WIS values (top) and empirical coverage rates (bottom) achieved by different models for the age-stratified SARI hospitalization incidence. Results are averaged over forecast dates and age groups, and stratified by nowcast/forecast horizon. The average scores are decomposed into components for overprediction, underprediction, and forecast spread.

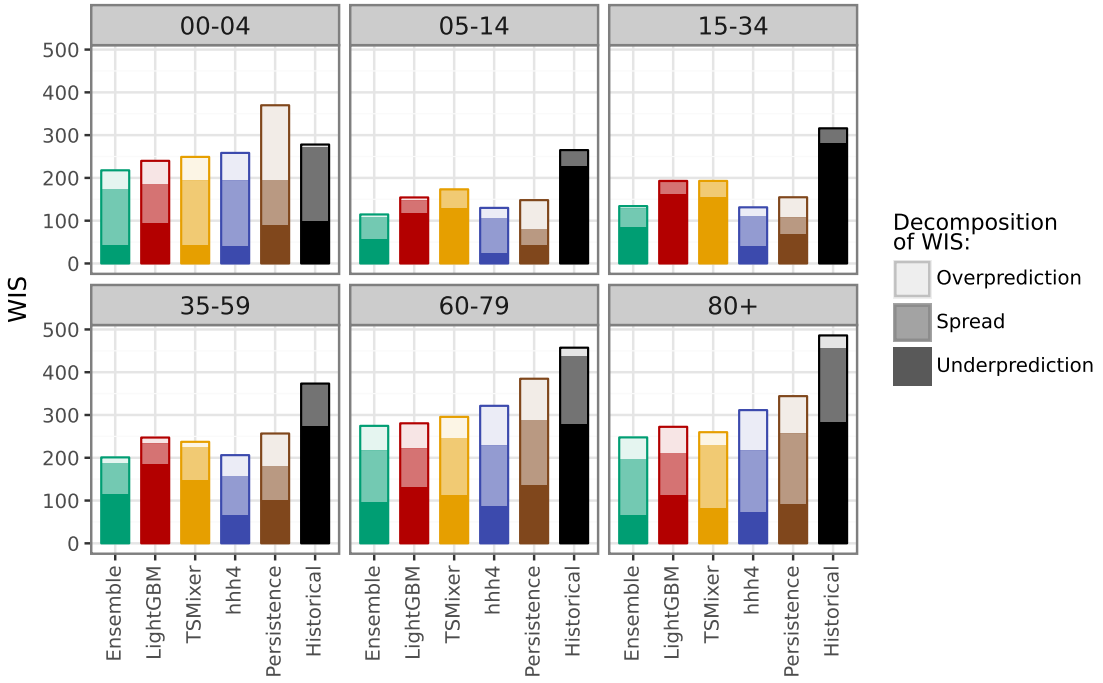


Figure 7.10.: Average WIS by age group, aggregated over forecast dates and horizons. Average scores are decomposed into components for overprediction, underprediction, and forecast spread.

The forecasts from the `LightGBM`, and to a lesser degree TSMixer models are somewhat overconfident, too. This is not surprising given that the forecasting models take the nowcast as an input. Remarkably, the `Ensemble` forecast is well-calibrated across horizons and interval levels. This can be explained by the fact that when using Vincentization, the length of the ensemble prediction intervals corresponds to the average length of the member intervals. If the ensemble intervals are centered around a more accurate central tendency (as is often the case), interval coverage rates will tend to increase.

Integration of now- and forecasts

For each of the forecasting methods, we investigate the impact of integrating nowcasts into forecasts and assesss the performance of the chosen implementation route. Thus, instead of including nowcast distributions in the way described in Section 7.3.2, we apply three alternative strategies.

(i) Firstly, we simply ignore the delay problem and use uncorrected incomplete data points to initialize our forecasting models ("Naive").

(ii) Secondly, we discard the last available observation and use only observations that are largely stable, as is common in the literature (Paireau et al., 2022). We still apply the nowcasting procedure to the previous weeks, but this makes little difference in practice ("Discard").

(iii) Lastly, we base forecasts on the final versions of the latest data points, i.e., assess how much forecasts would improve if the reporting system was free of delays. This is a hypothetical setting and not an approach that could be applied in real time ("Oracle").

Figure 7.11 summarizes the performance for the total SARI hospitalization incidence under the four considered ways of handling recent data points. Our proposed method of including the latest data point with a nowcast correction ("Coupling") yields improvements relative to using uncorrected data ("Naive") and discarding this data point ("Discard"). This holds especially for short horizons, where the initialization of forecasts is most relevant. In fact, for the `hhh4` model, the "Discard" version even works slightly better for horizons 3 and 4. Somewhat surprisingly, when providing forecast models with the final values of recent data points ("Oracle") rather than nowcasts, performance does not always improve. While for `hhh4` it does, for the other models there are minor deteriorations in performance for some horizons. A possible explanation is that initializing the models `LightGBM` and `TSMixer` with a nowcast distribution rather than the correct value increases forecast dispersion, which can lead to improved calibration. Corresponding results for age-stratified predictions are shown in Figure E.6 in the Appendix and are in good agreement with the aggregate-level results.

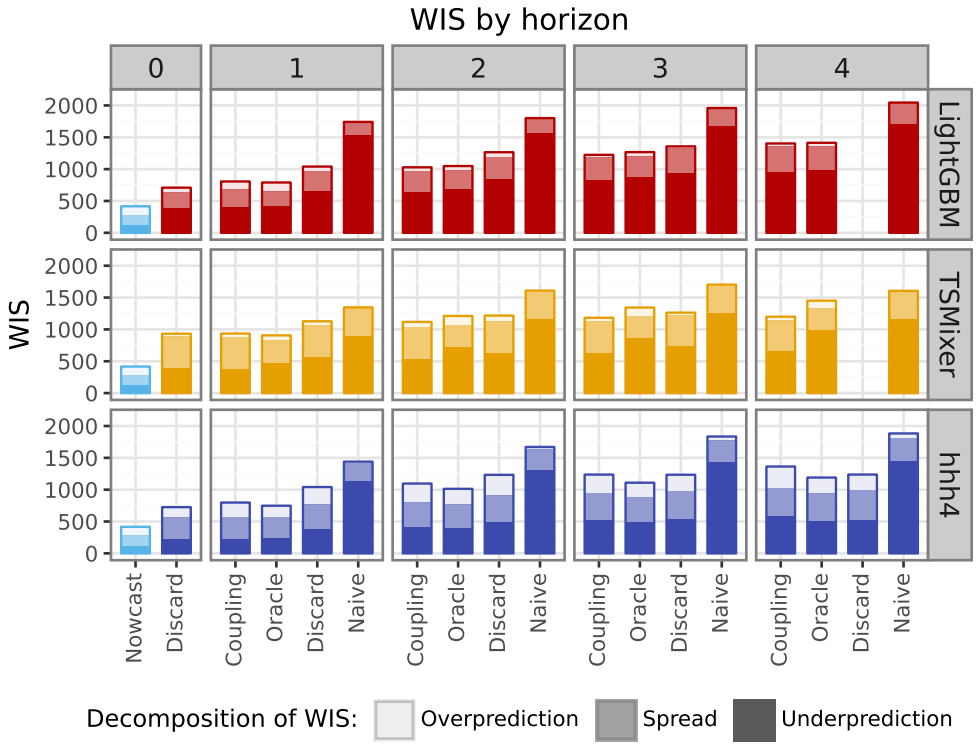


Figure 7.11.: Comparison of forecast performance on the aggregate level resulting from different strategies to handle incomplete recent data. "Coupling" is our main approach described in Section 7.3.2, i.e. feeding the full nowcast into forecasting models. "Discard" corresponds to discarding the most recent (i.e., most incomplete) data point and treating it like an additional value to be predicted. "Naive" uses the time series as is (with yet incomplete values). "Oracle" is a hypothetical setting where the final versions of the most recent data points are used. It thus enables us to assess the impact of reporting delays on forecast quality.

Robustness checks

Our primary approach consists of including all available data, that comprises the COVID-19 period for training and also uses additional information of ARI to predict SARI incidences. These choices were motivated by hyperparameter tuning, see Table E.1 in the Appendix, where the inclusion of the COVID-19 period and the use of additional ARI data were incorporated into the hyperparameter tuning process.

In this subsection, we study how these implementation decisions impact the predictive performance of each of the forecasting methods. In particular, we exclude the seasons affected by COVID-19 in the training data (see vertical dashed lines in Figure 7.2) and also omit additional ARI information. In each of the settings, the ML models were re-trained with the corresponding optimal hyperparameters. As before, due to the non-deterministic nature of the training process, we conducted training using 10 different random seeds and averaged the forecasts from these models (i.e., the predictive quantiles at each level) to obtain more robust results.

Figure E.7 in the Appendix summarizes the effect of excluding data from the COVID-19 period in the training set as well as from omitting ARI incidences as an auxiliary data stream for `LightGBM` and `TSMixer`. Using data from the COVID-19 period led to slight improvements in performance for `hhh4` and `LightGBM`. The `TSMixer` model was very poorly behaved when applied to a reduced data set without the COVID-19 period, indicating that our full time series may already be towards the lower end of the data requirements of this method. The inclusion of the auxiliary time series on outpatient consultations for ARI had only a minor impact on performance, yielding slight improvements for `TSMixer` and deteriorations for `LightGBM`.

7.5. Outlook: Prospective evaluation in the RESPINOW Hub

The presented work serves as a blueprint for a broader operational disease nowcasting and forecasting system called the *RESPINOW Hub*. Funded by the German Ministry of Education and Research (BMBF), it is conceived as an open and collaborative system accepting modeling results from multiple research groups. Both conceptually and concerning its technical implementation, it follows the conventions of the Forecast Hub ecosystem (Reich et al., 2022; Cramer et al., 2022b; Wolffram et al., 2023). While we here presented a retrospective forecasting exercise based on historical data snapshots, the *RESPINOW Hub* has been collecting real-time predictions on four prediction targets since fall 2024:

- the SARI hospitalization incidence as discussed in the present paper.
- the outpatient consultation incidence for acute respiratory infections (Goerlitz et al., 2021), which in the present work served as an auxiliary data stream.
- the incidence of laboratory-confirmed cases of seasonal influenza as well as respiratory syncytial virus (RSV) as published via the SURVSTAT@RKI 2.0 system (Robert Koch Institute, 2025).

All models presented in this work are also included in the *RESPINOW Hub*. They are complemented by additional nowcasting models and statistical forecasting approaches.This includes in particular a nowcasting approach developed by Robert Koch Institute which will be documented separately. Moreover, mechanistic models for individual pathogens like seasonal influenza are included. The resulting predictions are shared in a weekly rhythm via a dashboard (`http://respinowhub.de/`) as well as a public GitHub repository (`https://github.com/KITmetricslab/RESPINOW-Hub`). An evaluation study on the 2024/25 season has been preregistered (Bracher and Wolffram, 2024), and will shed light on the operational performance of the different models.

7.6. Discussion

We presented and evaluated a multi-model system for nowcasting and short-term forecasting of hospitalizations from severe acute respiratory infections (SARI) in Germany. We addressed this in a modular fashion, where nowcasts were generated in a separate step and subsequently fed into the forecasting models. For short forecast horizons, this led to clear improvements relative to a simpler approach where the most recent data points were used in an uncorrected fashion or simply discarded. Similarly to previous efforts, we found that combined ensemble predictions performed consistently better than individual forecasting models. Forecasts were generally well-calibrated in terms of interval coverage fractions, but in some models as well as the ensemble we observed noteworthy biases in some age groups. Especially the machine learning models `LightGBM` and `TSMixer` in these instances seemed to overfit to historical patterns, while the simpler statistical approach `hhh4` fared better. In age groups where the seasonal course was closer to historical patterns, however, this model had weaker relative performance.

The good probabilistic calibration of almost all considered models represents a marked difference from results achieved in recent years for COVID-19 cases or deaths (see e.g., Bracher et al. 2021b; Cramer et al. 2022b). This is surely not due to a sudden improvement in forecasting capacities, but due to the higher predictability of seasonal disease dynamics. Unlike in COVID-19 forecasting, social dynamics and intervention measures were likely no major drivers during the test period. Also, reporting practices were considerably more stable than for most COVID-19 indicators.

Our analyses of forecast performance across horizons and age groups indicate that our three stand-alone models have differing strengths and weaknesses. This *ensemble diversity* is often seen as a key feature for good ensemble performance (DelSole et al., 2014). Especially during the COVID-19 pandemic, collaborative forecasting projects featured considerably more models (the largest effort likely being Cramer et al. 2022b with more than 100 models). This level of effort is unrealistic (and undesirable) outside of times of major crisis. How many models need to be run in order to achieve robust ensemble performance is subject to current research. Fox et al. (2024) recommend using four to seven models and find that the gain from additional models diminishes quickly. In the *RESPINOW Hub*, two more independently run models have recently been included

for SARI. We hope this will further enhance the robustness of the ensemble, all while keeping the required effort at a sustainable level.

The ongoing *RESPINOW Hub* project will also enable us to address one of the major weaknesses of the present project, which is the risk of hindsight bias. While we made considerable efforts to manage historical data versions correctly and avoid using data that would not have been available in real time, the applied development and evaluation of prediction models is an iterative process. Implicitly some knowledge on characteristics of the test set may thus have diffused into our forecasting approaches. The follow-up evaluation study of real-time forecasts, which we preregistered (Bracher and Wolffram, 2024), will enable us to evaluate the developed models without the risk of hindsight bias.

In the present work, we were exclusively concerned with aggregate SARI hospitalizations which are unspecific to the causative agent. For the 2024/25 season, the Robert Koch Institute started releasing stratified data on SARI hospitalizations caused by COVID-19, seasonal influenza, and RSV. These represent highly relevant additional prediction targets and open new avenues for more mechanistic models explicitly reflecting the dynamics of infection and susceptibility. Such a stratified approach may ultimately also lead to improved forecasts of the total SARI hospitalization incidence.

Appendix E

E.1. Details on hyperparameter tuning

Table E.1.: Hyperparameter spaces for tuning

LightGBM

Parameter	Values
colsample_bytree	{0.8}
learning_rate	{0.01, 0.05, 0.1}
max_bin	{1024, 2048}
max_depth	{-1}
min_child_samples	{10, 20, 40}
min_split_gain	{0}
n_estimators	{500, 1000}
num_leaves	{20, 31, 40}
reg_alpha	{0, 0.5, 1}
reg_lambda	{0, 0.5, 1}
sample_weight	{linear, no-covid}
subsample	{0.8}
subsample_freq	{1}
use_covariates	{true, false}
Combinations:	3888
Runtime:	3 days

TSMixer

Parameter	Values
activation	{ReLU}
batch_size	{32}
dropout	{0.2}
ff_size	{32, 64}
hidden_size	{32, 64}
n_epochs	{500, 1000}
norm_type	{TimeBatchNorm2d}
normalize_before	{false}
num_blocks	{4, 6}
optimizer	{AdamW}
lr	{0.0005, 0.001, 0.005}
weight_decay	{0, 0.001, 0.0001}
sample_weight	{linear, no-covid}
use_covariates	{true, false}
Combinations:	576
Runtime:	6 days

Table E.2.: Optimal `LightGBM` hyperparameters for different settings, determined by tuning on the validation period.

LightGBM

Parameter	All Data	No Covariates	No Covid
colsample_bytree	0.8	0.8	0.8
learning_rate	0.1	0.01	0.1
lags	8	8	8
lags_future_covariates	[0, 1]	[0, 1]	[0, 1]
max_bin	1024	2048	2048
max_depth	-1	-1	-1
min_child_samples	10	40	10
min_split_gain	0	0	0
n_estimators	500	1000	500
num_leaves	40	40	31
reg_alpha	0.5	0	0.5
reg_lambda	0.5	0.5	0.5
sample_weight	linear	linear	no-covid
subsample	0.8	0.8	0.8
subsample_freq	1	1	1
use_covariates	true	false	true
WIS (validation)	447.06	453.60	448.50

Table E.3.: Optimal `TSMixer` hyperparameters for different settings, determined by tuning on the validation period.

TSMixer

Parameter	All Data	No Covariates	No Covid
activation	ReLU	ReLU	ReLU
batch_size	32	32	32
dropout	0.2	0.2	0.2
ff_size	64	64	32
hidden_size	32	64	32
input_chunk_length	8	8	8
normalize_before	false	false	false
norm_type	TimeBatchNorm2d	TimeBatchNorm2d	TimeBatchNorm2d
num_blocks	4	4	4
n_epochs	500	1000	500
optimizer	AdamW	AdamW	AdamW
lr	0.001	0.0005	0.005
weight_decay	0.001	0.0001	0.0001
subsample_freq	1	1	1
use_covariates	true	false	false
WIS (validation)	359.20	356.92	432.69

E.2. The ARI data set

The *Arbeitsgemeinschaft Influenza* sentinel surveillance system (Goerlitz et al., 2021) consists of more than 600 general practitioners, who voluntarily provide information on the number of consultations for respiratory infections. Reporting is done directly to RKI either electronically (SEED-ARE) or by fax. We use the consultation incidence for acute respiratory infections (ARI; ICD-10 codes J00 – J22, B34.9 and J44.0) per 100,000 inhabitants. This indicator is not specific to one pathogen and thus forms part of syndromic surveillance. Data are in principle available per age group and region (with certain pairs of German states merged).

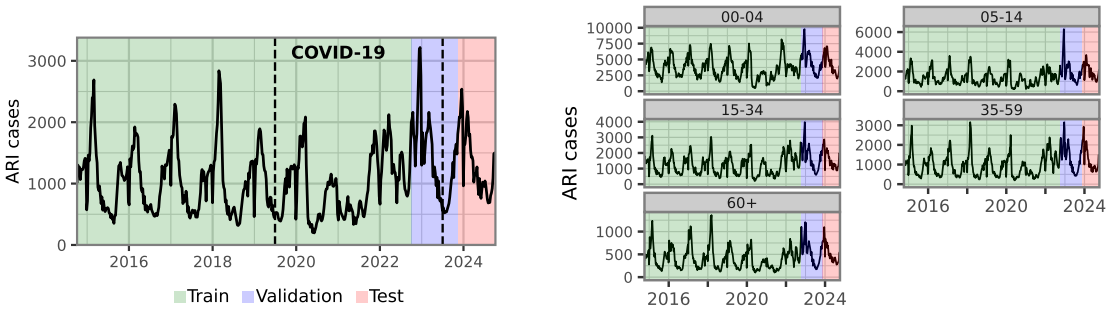


Figure E.1.: Time series of weekly ARI cases in Germany, 2014–2024. Colors indicate the split of the data into training, validation, and test data. The portion labeled "COVID-19" is only included in the training set for part of our model specifications.

E.3. Supplementary figures

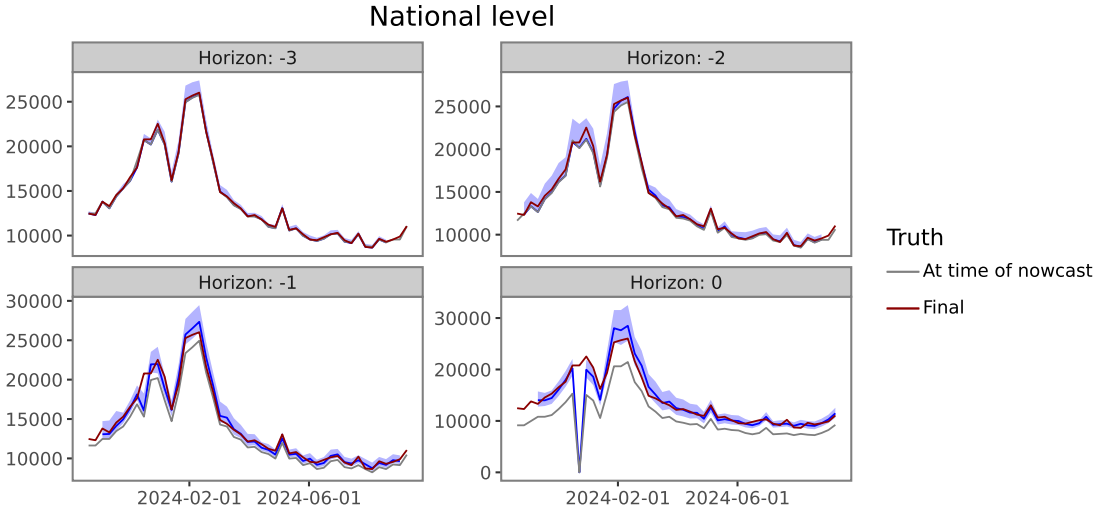


Figure E.2.: Nowcasts (blue) for the total SARI hospitalization incidence (pooled across age groups), stratified by the horizon. The final values (after 4 weeks) are shown in red, while the gray line represents the incomplete values available at the time of prediction.

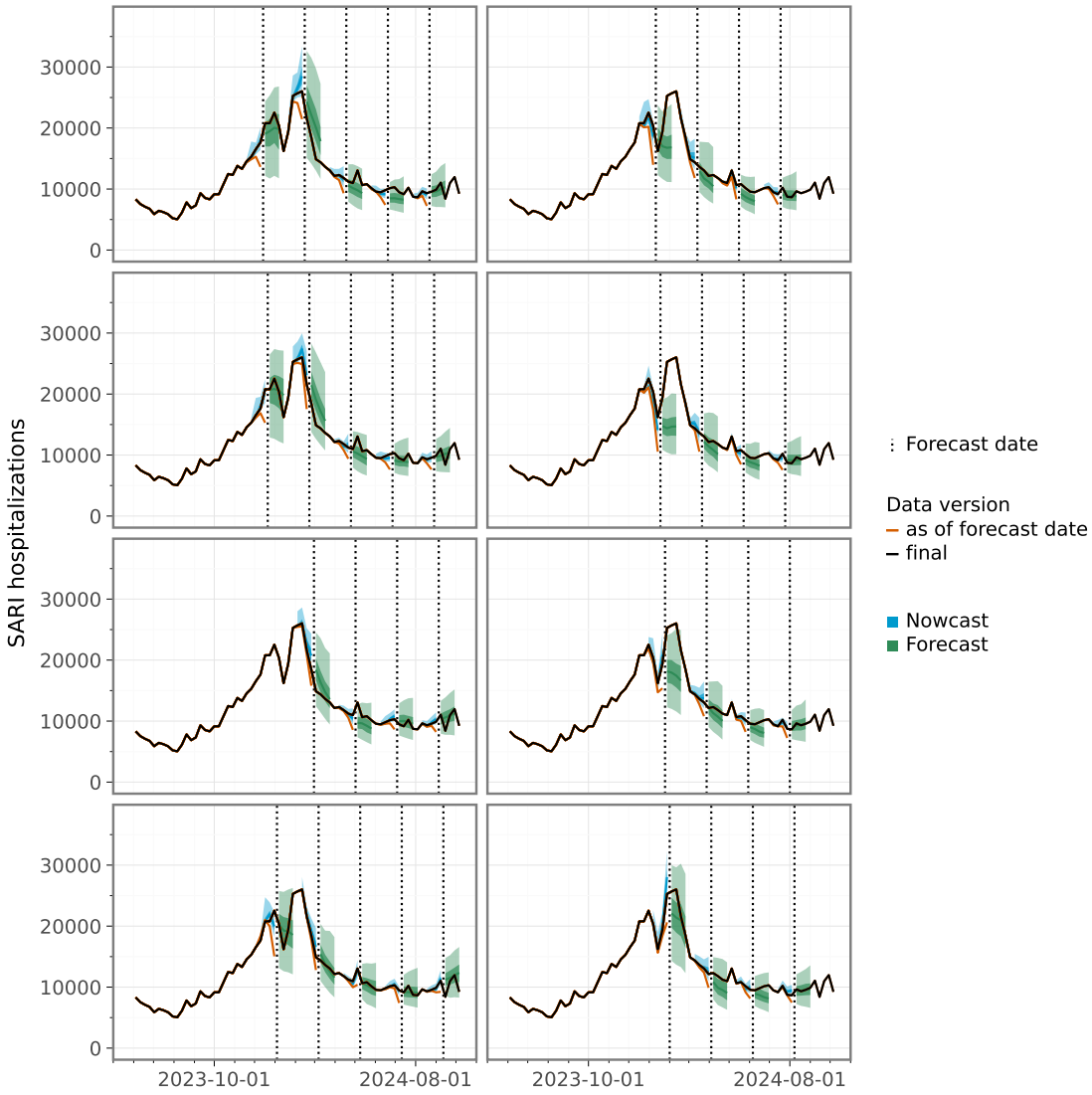


Figure E.3.: Nowcasts and ensemble forecasts for the total SARI hospitalization incidence (pooled across age groups) at different forecast times. To avoid overplotting, the predictions are displayed in multiple panels.

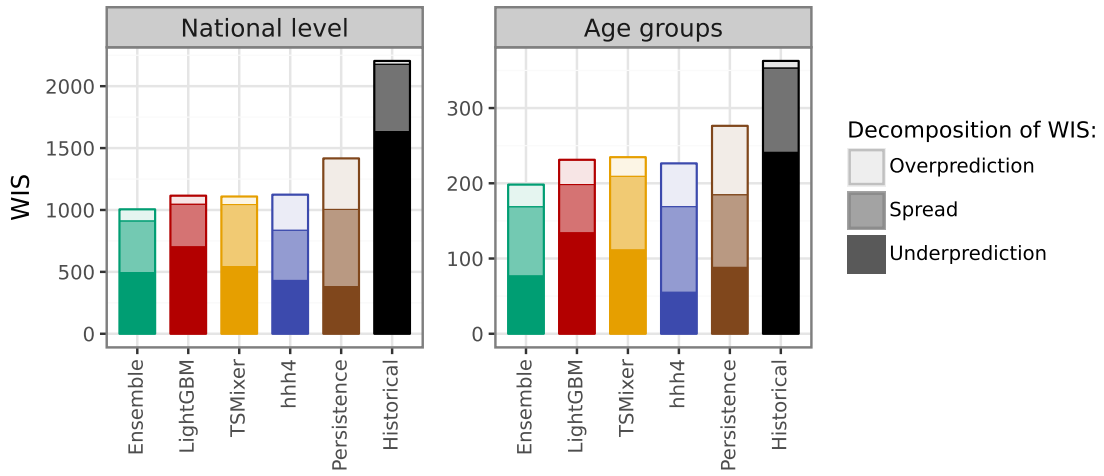


Figure E.4.: Average WIS on the national level and across age groups. Average scores are decomposed into components for overprediction, underprediction, and forecast spread.

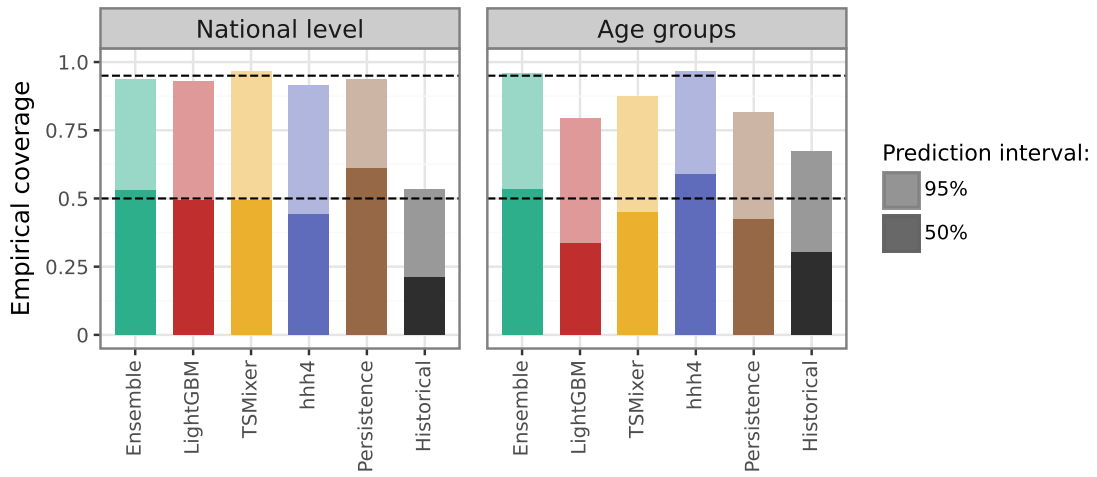


Figure E.5.: Empirical coverage of the central 50%- and 95%- prediction intervals of different models.

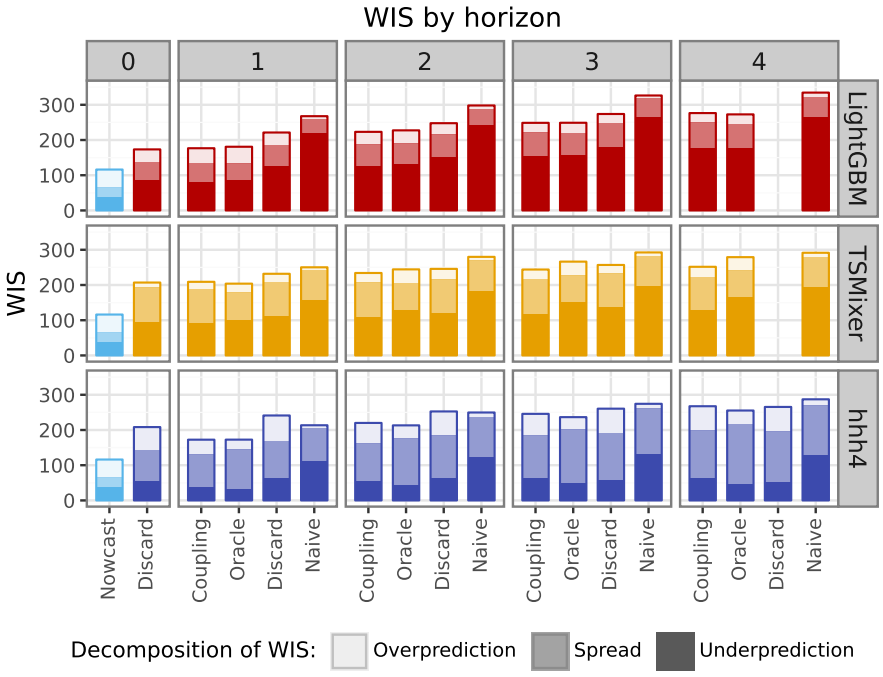


Figure E.6.: Comparison of forecast performance across age groups resulting from different strategies to handle incomplete recent data. "Coupling" is our main approach described in Section 7.3.2, i.e. feeding the full nowcast into forecasting models. "Discard" corresponds to discarding the most recent (i.e., most incomplete) data point and treating it like an additional value to be predicted. "Naive" uses the time series as is (with yet incomplete values). "Oracle" is a hypothetical setting where the final versions of the most recent data points are used. It thus enables us to assess the impact of reporting delays on forecast quality.

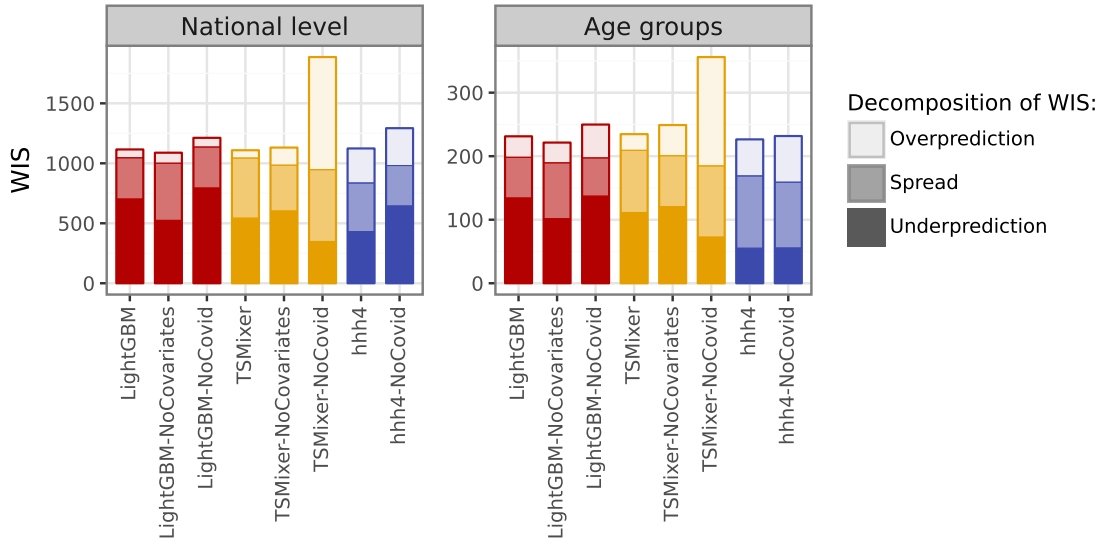


Figure E.7.: Average WIS (pooled across forecast dates and horizons) for different choices on the training data set. "NoCovid" denotes models where seasons strongly affected by the COVID-19 pandemic (see Figure 7.2 are excluded). "NoCovariates" denotes models where the auxiliary data set on ARI consultations was removed (note that the `hhh4` model does not use this in the first place).

Bibliography

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

Abbott, S., J. Hellewell, R. Thompson, K. Sherratt, H. Gibbs, N. Bosse, J. Munday, S. Meakin, E. Doughty, J. Chun, Y. Chan, F. Finger, P. Campbell, A. Endo, C. Pearson, A. Gimma, T. Russell, S. Flasche, A. Kucharski, R. Eggo, and S. Funk (2020b): "Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts," *Wellcome Open Research*, 5.

Abbott, S., A. Lison, S. Funk, C. Pearson, and H. Gruson (2021): "epinowcast: Flexible hierarchical nowcasting," Available from: https://github.com/epinowcast/epinowcast, doi: 10.5281/zenodo.5637165.

Adamik, B., M. Bawiec, V. Bezborodov, W. Bock, M. Bodych, J. P. Burgard, T. Götz, T. Krueger, A. Migalska, B. Pabjan, T. Ożański, E. Rafajłowicz, W. Rafajłowicz, E. Skubalska-Rafajłowicz, S. Ryfczyńska, E. Szczurek, and P. Szymański (2020): "Mitigation and herd immunity strategy for COVID-19 is likely to fail," *medRxiv*.

Adrian, T., N. Boyarchenko, and D. Giannone (2019): "Vulnerable growth," *American Economic Review*, 109, 1263–1289.

Afelt, A., R. Bartczuk, P. Biecek, M. Bodych, A. Gambin, K. Gogolewski, A. Kaczorek, J. Kisielewski, T. Krüger, Migalska, R. Mikołajczyk,

A. Moszyński, K. Niedzielewski, A. Nowosielski, B. Pabjan, M. Radwan, F. Rakowski, M. Rosińska, M. Semeniuk, and J. Zieliński (2020): "Quo vadis coronavirus? Rekomendacje zespołów epidemiologii obliczeniowej na rok 2021," Available online at https://quovadis.crs19.pl/ (last accessed 22 December 2020).

Amaral, A. V. R., D. Wolffram, P. Moraga, and J. Bracher (2024): "Post-processing and weighted combination of infectious disease nowcasts," *medRxiv*.

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

Arjovsky, M., S. Chintala, and L. Bottou (2017): "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, PMLR, 214–223.

Arik, S., J. Shor, R. Sinha, J. Yoon, J. Ledsam, L. Le, M. Dusenberry, N. Yoder, K. Popendorf, A. Epshteyn, J. Euphrosine, E. Kanal, I. Jones, C. Li, B. Luan, J. Mckenna, V. Menon, S. Singh, M. Sun, A. Ravi, L. Zhang, D. Sava, K. Cunningham, H. Kayama, T. Tsai, D. Yoneoka, S. Nomura, H. Miyata, and T. Pfister (2021): "A prospective evaluation of AI-augmented epidemiology to forecast COVID-19 in the USA and Japan," *npj Digital Medicine*, 4, 146.

Ayer, M., H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman (1955): "An empirical distribution function for sampling with incomplete information," *The Annals of Mathematical Statistics*, 641–647.

Baker, R. E., J.-M. Peña, J. Jayamohan, and A. Jérusalem (2018): "Mechanistic models versus machine learning, a fight worth fighting for the biological community?" *Biology Letters*, 14, 20170660.

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

Barna, D. M., K. Engeland, T. L. Thorarinsdottir, and C.-Y. Xu (2023): "Flexible and consistent Flood–Duration–Frequency modeling: A Bayesian approach," *Journal of Hydrology*, 620, 129448.

Basle Committee on Banking Supervision (1996): "Overview of the amendment to the capital accord to incorporate market risks," Tech. rep., Bank for International Settlements.

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

Becker, C., P. Duersch, and T. Eife (2023): "Measuring inflation expectations: How the response scale shapes density forecasts," *Preprint*, https://archiv.ub.uni-heidelberg.de/volltextserver/32677/1/Becker_Duersch_Eife_2023_dp723.pdf.

Beesley, L. J., D. Osthus, and S. Y. Del Valle (2022): "Addressing delayed case reporting in infectious disease forecast modeling," *PLOS Computational Biology*, 18, 1–26.

Bellemare, M. G., I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos (2017): "The Cramer distance as a solution to biased Wasserstein gradients," *arXiv preprint arXiv:1705.10743*.

Bentzien, S. and P. Friederichs (2014): "Decomposition and graphical portrayal of the quantile score," *Quarterly Journal of the Royal Meteorological Society*, 140, 1924–1934.

Bergström, F., F. Günther, M. Höhle, and T. Britton (2022): "Bayesian nowcasting with leading indicators applied to COVID-19 fatalities in Sweden," *PLOS Computational Biology*, 18, 1–17.

Berliner Morgenpost (2021): "Triage in Sachsen: Kliniken bereiten sich auf Schlimmes vor," Available from: https://web.archive.org/web/2024122316

3027/https://www.morgenpost.de/vermischtes/article233915811/corona-sachsen-triage-intensivstationen-ueberlastung.html, cited 19 July 2023.

Berndt, C., C. Endt, and S. Müller-Hansen (2021a): "Die unsichtbare Welle," *Süddeutsche Zeitung*, published online, 5 February 2021, https://web.archive.org/web/20241204063131/https://www.sueddeutsche.de/wissen/coronavirus-mutante-b117-daten-1.5197700.

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

Bhatia, S., K. V. Parag, J. Wardle, N. Imai, S. L. Van Elsland, B. Lassmann, G. Cuomo-Dannenburg, E. Jauneikaite, H. J. T. Unwin, S. Riley, N. Ferguson, C. A. Donnelly, A. Cori, and P. Nouvellet (2021): "Global predictions of short- to medium-term COVID-19 transmission trends : a retrospective assessment," *medRxiv*.

Bicher, M., M. Zuba, L. Rainer, F. Bachner, C. Rippinger, H. Ostermann, N. Popper, S. Thurner, and P. Klimek (2020): "Supporting Austria through the COVID-19 Epidemics with a Forecast-Based Early Warning System," *medRxiv*.

Biewald, L. (2020): "Experiment Tracking with Weights and Biases," Software available from wandb.com.

Bleichrodt, A., R. Luo, A. Kirpich, and G. Chowell (2024): "Evaluating the forecasting performance of ensemble sub-epidemic frameworks and other time series models for the 2022–2023 mpox epidemic," *Royal Society Open Science*, 11, 240248.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-nonparametric IV estimation of shape-invariant Engel curves," *Econometrica*, 75, 1613–1669.

Borchering, R. K., C. Viboud, E. Howerton, C. P. Smith, S. Truelove, and M. C. Runge (2021): "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.

Bosse, N. (2020): "epiforecasts/crowdforecastr: beta release," Available online at https://doi.org/10.5281/zenodo.4618520.

Bosse, N. I., S. Abbott, J. Bracher, H. Hain, B. J. Quilty, M. Jit, C. for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, E. van Leeuwen, A. Cori, and S. Funk (2021): "Comparing human and model-based forecasts of COVID-19 in Germany and Poland," *medRxiv*.

Bracher, J. and L. Held (2022): "Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction," *International Journal of Forecasting*, 38, 1221–1233.

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

Bracher, J., the German and Polish COVID-19 Forecast Hub Team, and Participants (2020): "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.

Bracher, J. and D. Wolffram (2024): "Preregistration: Nowcasting and Short-Term Forecasting of Respiratory Infections in Germany, 2024/25," https://osf.io/tgsem/.

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

Bracher, J., D. Wolffram, and the COVID-19 Nowcast Hub team and participants (2021c): "Study Protocol: Comparison and combination of COVID-19 hospitalization nowcasts in Germany," Registry of the Open Science Foundation, https://osf.io/mru75/, deposited 23 November 2021.

Bracher, J., D. Wolffram, the German, and P. C.-. F. H. Team (2022): "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).

Brehmer, J. and T. Gneiting (2021): "Scoring interval forecasts: Equal-tailed, shortest, and modal interval," *Bernoulli*, 27, 1993–2010.

Brehmer, J. and K. Strokorb (2019): "Why scoring functions cannot assess tail properties," *Electronic Journal of Statistics*, 13, 4015 – 4034.

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

Bröcker, J. (2009): "Reliability, sufficiency, and the decomposition of proper scores," *Quarterly Journal of the Royal Meteorological Society*, 135, 1512–1519.

Buchholz, U., A.-S. Lehfeld, K. Tolksdorf, W. Cai, J. Reiche, B. Biere, R. Dürrwald, and S. Buda (2023): "Respiratory infections in children and adolescents in Germany during the COVID-19 pandemic," *Journal of Health Monitoring*, 8, 20.

Buda, S., K. Tolksdorf, E. Schuler, R. Kuhlen, and W. Haas (2017): "Establishing an ICD-10 code based SARI-surveillance in Germany–description of the system and first results from five recent influenza seasons," *BMC Public Health*, 17, 1–13.

Burgard, J. P., S. Heyder, T. Hotz, and T. Krueger (2021): "Regional estimates of reproduction numbers with application to COVID-19," *arXiv preprint arXiv:2108.13842*.

Busetti, F. (2017): "Quantile Aggregation of Density Forecasts," *Oxford Bulletin of Economics and Statistics*, 79, 495–512.

Castro, L., G. Fairchild, I. Michaud, and D. Osthus (2021): "COFFEE: COVID-19 Forecasts using Fast Evaluations and Estimation," *arXiv preprint arXiv:2110.01546*.

Castro, M., S. Ares, J. Cuesta, and S. Manrubia (2020): "The turning point and end of an expanding epidemic cannot be precisely forecast," *Proceedings of the National Academy of Sciences*, 117, 26190–26196.

Chen, S.-A., C.-L. Li, N. Yoder, S. O. Arik, and T. Pfister (2023a): "TSMixer: An All-MLP Architecture for Time Series Forecasting," *arXiv preprint arXiv:2303.06053*.

Chen, Y., Z. Lin, and H.-G. Müller (2023b): "Wasserstein regression," *Journal of the American Statistical Association*, 118, 869–882.

Chen, Z., A. Gaba, I. Tsetlin, and R. L. Winkler (2022): "Evaluating quantile forecasts in the M5 uncertainty competition," *International Journal of Forecasting*, 38, 1531–1545.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010): "Quantile and probability curves without crossing," *Econometrica*, 78, 1093–1125.

Choe, Y. and A. Ramdas (2021): "Comparing sequential forecasters," *arXiv preprint arXiv:2110.00115*.

Christoffersen, P. F. (1998): "Evaluating interval forecasts," *International Economic Review*, 39, 841–862.

Chung, Y., W. Neiswanger, I. Char, and J. Schneider (2021): "Beyond pinball loss: Quantile methods for calibrated uncertainty quantification," *Advances in Neural Information Processing Systems*, 34, 10971–10984.

Coibion, O. and Y. Gorodnichenko (2012): "What Can Survey Forecasts Tell Us about Information Rigidities?" *Journal of Political Economy*, 120, 116–159.

Coibion, O., Y. Gorodnichenko, and M. Weber (2022): "Monetary policy communications and their effects on household inflation expectations," *Journal of Political Economy*, 130, 1537–1584.

Conde-Amboage, M., I. Van Keilegom, and W. González-Manteiga (2021): "A new lack-of-fit test for quantile regression with censored data," *Scandinavian Journal of Statistics*, 48, 655–688.

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

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

Cramer, E. Y., Y. Huang, Y. Wang, E. L. Ray, M. Cornell, J. Bracher, A. Brennen, et al. (2022a): "The United States COVID-19 Forecast Hub Dataset," *Scientific Data*, 9.

Cramer, E. Y., E. L. Ray, V. K. Lopez, J. Bracher, A. Brennen, A. J. Castro Rivadeneira, A. Gerding, T. Gneiting, K. H. House, Y. Huang, et al. (2022b): "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States," *Proceedings of the National Academy of Sciences*, 119, e2113561119.

Czado, C., T. Gneiting, and L. Held (2009): "Predictive model assessment for count data," *Biometrics*, 65, 1254–61.

Davies, N., S. Abbott, R. Barnard, C. Jarvis, A. Kucharski, J. Munday, C. Pearson, T. Russell, D. Tully, A. Washburne, T. Wenseleers, GA, W. Waites, K. Wong, K. van Zandvoort, J. Silverman, CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, K. Diaz-Ordaz, R. Keogh, R. Eggo, S. Funk, M. Jit, K. Atkins, and W. Edmunds

(2021): "Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England," *Science*, 372, eabg3055.

Dawid, A. P. (1984): "Statistical theory: The prequential approach," *Journal of the Royal Statistical Society: Series A (General)*, 147, 278–290.

De Backer, M., A. E. Ghouch, and I. Van Keilegom (2019): "An adapted loss function for censored quantile regression," *Journal of the American Statistical Association*, 114, 1126–1137.

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

del Barrio, E., J. A. Cuesta-Albertos, C. Matran, and J. M. Rodriguez-Rodriguez (1999): "Tests of goodness of fit based on the L2-Wasserstein distance," *The Annals of Statistics*, 27, 1230–1239.

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

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

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

Deza, M. and E. Deza (2013): *Encyclopedia of Distances*, Springer Berlin, Heidelberg.

Diebold, F. X. and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, 13, 253–63.

Dimitriadis, T., T. Gneiting, and A. I. Jordan (2021): "Stable reliability diagrams for probabilistic classifiers," *Proceedings of the National Academy of Sciences*, 118, e2016191118.

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

Donat, M. G., L. V. Alexander, H. Yang, I. Durre, R. Vose, R. J. H. Dunn, K. M. Willett, E. Aguilar, M. Brunet, J. Caesar, B. Hewitson, C. Jack, A. M. G. Klein Tank, A. C. Kruger, J. Marengo, T. C. Peterson, M. Renom, C. Oria Rojas, M. Rusticucci, J. Salinger, A. S. Elrayah, S. S. Sekele, A. K. Srivastava, B. Trewin, C. Villarroel, L. A. Vincent, P. Zhai, X. Zhang, and S. Kitching (2013): "Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset," *Journal of Geophysical Research: Atmospheres*, 118, 2098–2118.

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

Donker, T., M. van Boven, W. van Ballegooijen, T. van't Klooster, C. Wielders, and J. Wallinga (2011): "Nowcasting pandemic influenza A/H1N1 2009 hospitalizations in the Netherlands," *European Journal of Epidemiology*, 26, 195–201.

Duffie, D. and J. Pan (1997): "An overview of value at risk," *Journal of Derivatives*, 4, 7–49.

Ehm, W., T. Gneiting, A. Jordan, and F. Krüger (2016): "Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 505–562.

Engel, E. (1857): "Die vorherrschenden Gewerbszweige in den Gerichtsämtern mit Beziehung auf die Productions- und Consumtionsverhältnisse des Königreichs Sachsen,"

Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministeriums des Innern, 8–9, 153–82.

Engelberg, J., C. F. Manski, and J. Williams (2009): "Comparing the point predictions and subjective probability distributions of professional forecasters," *Journal of Business & Economic Statistics*, 27, 30–41.

England, P. D. and R. J. Verrall (2002): "Stochastic claims reserving in general insurance," *British Actuarial Journal*, 8, 443–518.

European Centre for Disease Prevention and Control (2020a): "Baseline projections of COVID-19 in the EU/EEA and the UK: Update." Published 17 September 2020, https://web.archive.org/web/20250131100036/https://www.ecdc.europa.eu/sites/default/files/documents/ECDC-30-day-projections-Sept-2020.pdf.

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

——— (2020c): "Projected baselines of COVID-19 in the EU/EEA and the UK for assessing the impact of de-escalation of measures," Published 26 May 2020, https://web.archive.org/web/20250131100836/https://www.ecdc.europa.eu/sites/default/files/documents/Projected-baselines-COVID-19-for-assessing-impact-measures.pdf.

Fasiolo, M., S. N. Wood, M. Zaffran, R. Nedellec, and Y. Goude (2021): "Fast calibrated additive quantile regression," *Journal of the American Statistical Association*, 116, 1402–1412.

Fischer-Fels, J. (2021): "Erste Hochrechnung zur Verbreitung der Coronamutationen," *Ärzteblatt*, published online, 3 February 2021, https://web.archive.org/web/20250131094313/https://www.aerzteblatt.de/nachrichten/120768/Erste-Hochrechnung-zur-Verbreitung-der-Corona-Mutationen.

Fissler, T., R. Frongillo, J. Hlavinov'a, and B. Rudloff (2021): "Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals," *Electronic Journal of Statistics*, 15, 1034–1084.

Fissler, T. and S. M. Pesenti (2023): "Sensitivity measures based on scoring functions," *European Journal of Operational Research*, 307, 1408–1423.

Fox, S. J., M. Kim, L. A. Meyers, N. G. Reich, and E. L. Ray (2024): "Optimizing disease outbreak forecast ensembles," *Emerging Infectious Diseases*, 30, 1967.

Fritz, C., G. De Nicola, M. Rave, M. Weigert, Y. Khazaei, U. Berger, H. Küchenhoff, and G. Kauermann (2022): "Statistical modelling of COVID-19 data: Putting generalized additive models to work," *Statistical Modelling*, 0.

Frogner, C., C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio (2015): "Learning with a Wasserstein loss," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, 2053–2061.

Fuhrmann, J. and M. Barbarossa (2020): "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.

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

Funk, S., A. Camacho, A. J. Kucharski, R. Lowe, R. M. Eggo, and W. J. Edmunds (2019): "Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014–15," *PLOS Computational Biology*, 15, e1006785.

Funk, S., E. Gilad, C. Watkins, and V. A. A. Jansen (2009): "The spread of awareness and its impact on epidemic outbreaks," *Proceedings of the National Academy of Sciences*, 106, 6872–6877.

Gandy, A., K. Jana, and A. E. Veraart (2022): "Scoring predictions at extreme quantiles," *AStA Advances in Statistical Analysis*, 106, 527–544.

Gasthaus, J., K. Benidis, Y. Wang, S. Rangapuram, and D. e. a. Salinas (2019): "Probabilistic forecasting with spline quantile function RNNs," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

Gelman, A., Y. Goegebeur, F. Tuerlinckx, and I. Van Mechelen (2000): "Diagnostic checks for discrete data regression models using posterior predictive simulations," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49, 247–268.

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

German and Polish COVID-19 Forecast Hub Team (2021a): "Codes to accompany Bracher et al: A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave," Available online at https://github.com/KITmetricslab/analyses_de_pl and https://doi.org/10.5281/zenodo.5085398.

——— (2021b): "German and Polish COVID-19 Forecast Hub," Available online at https://github.com/KITmetricslab/covid19-forecast-hub-de and https://doi.org/10.5281/zenodo.4752079 (stable release).

German Federal Government (2021): "Videoschaltkonferenz der Bundeskanzlerin mit den Regierungschefinnen und Regierungschefs der Länder am 18. November 2021," Available from: https://web.archive.org/web/20240627064504/https://www.bundesregierung.de/resource/blob/974430/1982598/defbdff47daf5f177586a5d34e8677e8/2021-11-18-mpk-data.pdf, cited 19 July 2023.

German Federal Ministry of Health (2021): "FAQ zur Hospitalisierungsinzidenz," Available from: https://web.archive.org/web/20250131093717/https://www.bundesgesundheitsministerium.de/coronavirus/hospitalisierungsinzidenz.html, cited 19 July 2023.

Giacomini, R. and I. Komunjer (2005): "Evaluation and combination of conditional quantile forecasts," *Journal of Business & Economic Statistics*, 23, 416–431.

Giannone, D., L. Reichlin, and D. Small (2008): "Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics*, 55, 665–676.

Gini, C. (1914): "Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazioni statistiche," *Atti Reale Inst. Veneto Sci. Lett. Arti*, 78, 185–213.

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

Gneiting, T. (2011a): "Making and evaluating point forecasts," *Journal of the American Statistical Association*, 106, 746–762.

——— (2011b): "Quantiles as optimal point forecasts," *International Journal of Forecasting*, 27, 197–207.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007): "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 243–268.

Gneiting, T. and M. Katzfuss (2014): "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, 1, 125–151.

Gneiting, T. and A. E. Raftery (2005): "Weather Forecasting with Ensemble Methods," *Science*, 310, 248–249.

——— (2007): "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102, 359–378.

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

Gneiting, T. and R. Ranjan (2011): "Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules," *Journal of Business & Economic Statistics*, 29, 411–422.

——— (2013): "Combining predictive distributions," *Electronic Journal of Statistics*, 7, 1747–1882.

Gneiting, T. and J. Resin (2023): "Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination," *Electronic Journal of Statistics*, 17, 3226–3286.

Goerlitz, L., K. Tolksdorf, U. Buchholz, K. Prahm, U. Preuss, M. an der Heiden, T. Wolff, R. Dürrwald, A. Nitsche, J. Michel, W. Haas, and S. Buda (2021): "Überwachung von COVID-19 durch Erweiterung der etablierten Surveillance für Atemwegsinfektionen," *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 64, 1437–1588, https://doi.org/10.1007/s00103-021-03303-2.

Golding, N., F. Shearer, R. Moss, P. Dawson, D. Liu, J. Ross, R. Hyndman, C. Zachreson, N. Geard, J. McVernon, D. Price, and J. McCaw (2020): "Estimating temporal variation in transmission of SARS-CoV-2 and physical distancing behaviour in Australia," Tech. rep., Doherty Institute, University of Melbourne, available online at https://web.archive.org/web/20211007225823/https://www.doherty.edu.au/uploads/content_doc/Technical_report_4_July17.pdf.

González Ordiano, J., L. Gröll, R. Mikut, and V. Hagenmeyer (2020): "Probabilistic energy forecasting using the nearest neighbors quantile filter and quantile regression," *International Journal of Forecasting*, 36, 310–323.

Grant, K. and T. Gneiting (2013): "Consistent scoring functions for quantiles," in *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner*, ed. by M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii, and M. Maathuis, Institute of Mathematical Statistics, 163–173.

Greene, S. K., S. F. McGough, G. M. Culp, L. E. Graf, M. Lipsitch, N. A. Menzies, and R. Kahn (2021): "Nowcasting for Real-Time COVID-19 Tracking in New York City: An Evaluation Using Reportable Disease Data From Early in the Pandemic," *JMIR Public Health Surveill*, 7, e25538.

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

Hale, T., N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, and H. Tatlow (2021): "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)," *Nature Human Behaviour*, 5, 529–538.

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

Hasell, J., E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino, E. Ortiz-Ospina, M. Roser, and H. Ritchie (2020): "A cross-country database of COVID-19 testing," *Scientific Data*, 7, 345.

Hawryluk, I., H. Hoeltgebaum, S. Mishra, X. Miscouridou, R. P. Schnekenberg, C. Whittaker, M. Vollmer, S. Flaxman, S. Bhatt, and T. A. Mellan (2021): "Gaussian process nowcasting: application to COVID-19 mortality reporting," in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, PMLR, vol. 161 of *Proceedings of Machine Learning Research*, 1258–1268.

Heinsch, M. and J. Schmid-Johannsen (2022): "Mit oder wegen Corona im Krankenhaus? So bedingt aussagekräftig sind die BW-Daten," Available from: https://web.archive.org/web/20230330061650/https://www.swr.de/swrak

tuell/baden-wuerttemberg/was-sagt-die-hospitalisierungsinzidenz-in-der-omikron-welle-aus-100.html, cited 19 July 2023.

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

Henzi, A. (2021): "isodistrreg: Isotonic Distributional Regression (IDR)," .

Henzi, A., J. Ziegel, and T. Gneiting (2021): "Isotonic distributional regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 963–993.

Hersbach, H. (2000): "Decomposition of the continuous ranked probability score for ensemble prediction systems," *Weather and Forecasting*, 15, 559–570.

Heyder, S. and T. Hotz (2023): "The ILM-prop model: method and code," Available from: https://github.com/Stochastik-TU-Ilmenau/ILM-prop.

Hoga, Y. and T. Dimitriadis (2022): "On testing equal conditional predictive ability under measurement error," *Journal of Business & Economic Statistics*.

Homburg, A., C. H. Weiss, L. C. Alwan, G. Frahm, and R. Göb (2019): "Evaluating approximate point forecasting of count processes," *Econometrics*, 7, 30.

Hong, T., P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. Hyndman (2016): "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond," *International Journal of Forecasting*, 32, 896–913.

Hong, T., P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour (2020): "Energy forecasting: A review and outlook," *IEEE Open Access Journal of Power and Energy*, 7, 376–388.

Howerton, E., L. Contamin, L. C. Mullany, M. Qin, N. G. Reich, S. Bents, R. K. Borchering, S.-m. Jung, S. L. Loo, C. P. Smith, et al. (2023): "Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty," *Nature Communications*, 14, 7260.

Huber, P. and E. Ronchetti (2009): *Robust Statistics*, Wiley, 2nd ed.

Hyndman, R. (2021): "forecast: Forecasting functions for time series and linear models," R package version 8.12.0, Available online: https://pkg.robjhyndman.com/forecast/.

Hyndman, R. and Y. Khandakar (2008): "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software*, 27, 1–22.

Höhle, M. and M. an der Heiden (2014): "Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011," *Biometrics*, 70, 993–1002.

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

Irpino, A. and R. Verde (2015): "Basic statistics for distributional symbolic variables: a new metric-based approach," *Advances in Data Analysis and Classification*, 9, 143–175.

Janke, T. and F. Steinke (2019): "Forecasting the price distribution of continuous intraday electricity trading," *Energies*, 12, 4262.

Jersakova, R., J. Lomax, J. Hetherington, B. Lehmann, G. Nicholson, M. Briers, and C. Holmes (2022): "Bayesian imputation of COVID-19 positive test counts for nowcasting under reporting lag," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71, 834–860.

Johansson, M. A., K. M. Apfeldorf, S. Dobson, J. Devita, A. L. Buczak, B. Baugher, L. J. Moniz, T. Bagley, S. M. Babin, E. Guven, T. K. Yamana, J. Shaman, T. Moschou, N. Lothian, A. Lane, G. Osborne, G. Jiang, L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, R. Rosenfeld, J. Lessler, N. G. Reich, D. A. T. Cummings, S. A. Lauer, S. M. Moore, H. E. Clapham, R. Lowe, T. C. Bailey, M. García-Díez, M. S. Carvalho, X. Rodó, T. Sardar, R. Paul, E. L. Ray, K. Sakrejda, A. C. Brown, X. Meng, O. Osoba, R. Vardavas, D. Manheim, M. Moore, D. M. Rao, T. C. Porco, S. Ackley, F. Liu, L. Worden, M. Convertino, Y. Liu, A. Reddy, E. Ortiz, J. Rivero,

H. Brito, A. Juarrero, L. R. Johnson, R. B. Gramacy, J. M. Cohen, E. A. Mordecai, C. C. Murdock, J. R. Rohr, S. J. Ryan, A. M. Stewart-Ibarra, D. P. Weikel, A. Jutla, R. Khan, M. Poultney, R. R. Colwell, B. Rivera-García, C. M. Barker, J. E. Bell, M. Biggerstaff, D. Swerdlow, L. Mier-y Teran-Romero, B. M. Forshey, J. Trtanj, J. Asher, M. Clay, H. S. Margolis, A. M. Hebbeler, D. George, and J.-P. Chretien (2019): "An open challenge to advance probabilistic forecasting for dengue epidemics," *Proceedings of the National Academy of Sciences*, 116, 24268–24274.

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

Jordan, A. (2016): "Facets of Forecast Evaluation," Ph.D. thesis, Karlsruhe Institute of Technology (KIT), https://publikationen.bibliothek.kit.edu/1000063629.

Jordan, A., A. Mühlemann, and J. Ziegel (2022): "Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals," *Annals of the Institute of Statistical Mathematics*, 74, 489–514.

Jose, V. and R. Winkler (2009): "Evaluating quantile assessments," *Operations Research*, 57, 1287–1297.

Karlen, D. (2020): "Characterizing the spread of CoViD-19," *arXiv preprint arXiv:2007.07156*.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017): "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, 30.

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

Keyel, A. C. and A. M. Kilpatrick (2021): "Probabilistic Evaluation of Null Models for West Nile Virus in the United States," *bioRxiv*.

Kheifetz, Y., H. Kirsten, and M. Scholz (2021): "On the parametrization of epidemiologic models – lessons from modelling COVID-19 epidemic," *arXiv preprint arXiv:2109.11916*.

Koenker, R. (2005): *Quantile regression*, vol. 38, Cambridge University Press.

——— (2017): "Quantile regression: 40 years on," *Annual Review of Economics*, 9, 155–176.

Koenker, R. and G. Bassett (1978): "Regression quantiles," *Econometrica*, 46, 33–50.

Koenker, R. and J. Machado (1999): "Goodness of fit and related inference processes for quantile regression," *Journal of the American Statistical Association*, 94, 1296–1310.

Krymova, E., B. Béjar, D. Thanou, T. Sun, E. Manetti, G. Lee, K. Namigai, C. Choirat, A. Flahault, and G. Obozinski (2022): "Trend estimation and short-term forecasting of COVID-19 cases and deaths worldwide," *Proceedings of the National Academy of Sciences*, 119, e2112656119.

Krüger, F. and J. Ziegel (2021): "Generic conditions for forecast dominance," *Journal of Business & Economic Statistics*, 39, 972–983.

Küchenhoff, H., F. Günther, G. Kauermann, W. Hartl, and G. Krause (2021): "Neuaufnahmen auf Intensivstationen als Alternative zu den Meldeinzidenzen als gesetzliche Grundlage für Maßnahmen zum Infektionsschutz," *CODAG Bericht Nr. 13, LMU München*, 2–6.

Landesgesundheitsamt Baden Württemberg (2021): "Tagesbericht COVID-19, Montag 8.2.2021," Available at https://web.archive.org/web/20250131093337/https://www.gesundheitsamt-bw.de/fileadmin/LGA/_DocumentLibraries/SiteCollectionDocuments/05_Service/LageberichtCOVID19/COVID_Lagebericht_LGA_210208.pdf.

Lawless, J. (1994): "Adjustments for reporting delays and the prediction of occurred but not reported events," *Canadian Journal of Statistics*, 22, 15–31.

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

Li, M. L., H. Tazi Bouardi, O. Skali Lami, T. A. Trikalinos, N. K. Trichakis, and D. Bertsimas (2020): "Forecasting COVID-19 and Analyzing the Effect of Government Interventions," *medRxiv*.

Li, R. and L. Peng (2017): "Assessing quantile prediction with censored quantile regression models," *Biometrics*, 73, 517–528.

Li, T. and L. F. White (2021): "Bayesian back-calculation and nowcasting for line list data during the COVID-19 pandemic," *PLOS Computational Biology*, 17, 1–22.

Lichtendahl, K. C., Y. Grushka-Cockayne, and R. L. Winkler (2013): "Is It Better to Average Probabilities or Quantiles?" *Management Science*, 59, 1594–1611.

Lorenzo, L. and J. Arroyo (2022): "Analysis of the cryptocurrency market using different prototype-based clustering techniques," *Financial Innovation*, 8, 7.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos (2022): "M5 accuracy competition: Results, findings, and conclusions," *International Journal of Forecasting*, 38, 1346–1364.

Manski, C. F. (2004): "Measuring expectations," *Econometrica*, 72, 1329–1376.

McDonald, D. J., J. Bien, A. Green, A. J. Hu, N. DeFries, S. Hyun, N. L. Oliveira, J. Sharpnack, J. Tang, R. Tibshirani, V. Ventura, L. Wasserman, and R. J. Tibshirani (2021): "Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction?" *Proceedings of the National Academy of Sciences*, 118, e2111453118.

McGough, S. F., M. A. Johansson, M. Lipsitch, and N. A. Menzies (2020): "Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking," *PLOS Computational Biology*, 16, e1007735.

McGowan, C. J., M. Biggerstaff, M. Johansson, K. M. Apfeldorf, M. Ben-Nun, L. Brooks, M. Convertino, M. Erraguntla, D. C. Farrow, J. Freeze, S. Ghosh, S. Hyun, S. Kandula, J. Lega, Y. Liu, N. Michaud, H. Morita, J. Niemi, N. Ramakrishnan, E. L. Ray, N. G. Reich, P. Riley, J. Shaman, R. Tibshirani, A. Vespignani, Q. Zhang, C. Reed, R. Rosenfeld, N. Ulloa, K. Will, J. Turtle, D. Bacon, S. Riley, W. Yang, and The Influenza Forecasting Working Group (2019): "Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016," *Scientific Reports*, 9, 683.

Meinshausen, N. (2006): "Quantile regression forests," *Journal of Machine Learning Research*, 7, 983–999.

Menkir, T. F., H. Cox, C. Poirier, M. Saul, S. Jones-Weekes, C. Clementson, P. M. de Salazar, M. Santillana, and C. O. Buckee (2021): "A nowcasting framework for correcting for reporting delays in malaria surveillance," *PLOS Computational Biology*, 17, e1009570.

Meyer, S., L. Held, and M. Höhle (2017): "Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance," *Journal of Statistical Software*, 77, 1–55.

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

Moran, K. R., G. Fairchild, N. Generous, K. Hickmann, D. Osthus, R. Priedhorsky, J. Hyman, and S. Y. Del Valle (2016): "Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast," *The Journal of Infectious Diseases*, 214, S404–S408.

Müller, A. and D. Stoyan (2002): *Comparison Methods for Stochastic Models and Risks*, Wiley, Chichester.

Munk, A. and C. Czado (1998): "Nonparametric validation of similar distributions and assessment of goodness of fit," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60, 223–241.

Murphy, A. and E. Epstein (1989): "Skill scores and correlation coefficients in model verification," *Monthly Weather Review*, 117, 572–581.

Mösching, A. and L. Dümbgen (2020): "Monotone least squares and isotonic quantiles," *Electronic Journal of Statistics*, 14, 24–49.

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

Ni, K., X. Bresson, T. Chan, and S. Esedoglu (2009): "Local histogram based segmentation using the Wasserstein distance," *International Journal of Computer Vision*, 84, 97–111.

Noh, H., A. El Ghouch, and I. Van Keilegom (2013): "Assessing model adequacy in possibly misspecified quantile regression," *Computational Statistics & Data Analysis*, 57, 558–569.

Nolde, N. and J. Ziegel (2017): "Elicitability and backtesting: Perspectives for banking regulation," *Annals of Applied Statistics*, 11, 1833–1874.

Norddeutscher Rundfunk (2021): "Nach MPK-Beschluss: Verwirrung um Hospitalisierungsinzidenz," Available from: https://web.archive.org/web/20230813155040/https://www.ndr.de/nachrichten/info/Nach-MPK-Beschluss-Verwirrung-um-Hospitalisierungsinzidenz,hospitalisierungsinzidenz100.html, cited 19 July 2023.

Orjebin, E. (2014): "A recursive formula for the moments of a truncated univariate normal distribution," *Preprint*, https://people.smp.uq.edu.au/YoniNazarathy/teaching_projects/studentWork/EricOrjebin_TruncatedNormalMoments.pdf.

Paireau, J., A. Andronico, N. Hozé, M. Layan, P. Crépey, A. Roumagnac, M. Lavielle, P.-Y. Boëlle, and S. Cauchemez (2022): "An ensemble model based

on early predictors to forecast COVID-19 health care demand in France," *Proceedings of the National Academy of Sciences*, 119, e2103302119.

Panaretos, V. M. and Y. Zemel (2019): "Statistical aspects of Wasserstein distances," *Annual Review of Statistics and Its Application*, 6, 405–431.

Patton, A. (2011): "Volatility forecast comparison using imperfect volatility proxies," *Journal of Econometrics*, 160, 246–256.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, and . B. e. a. Thirion (2011): "Scikit-learn: Machine learning in Python," in *Proceedings of the 12th Python Scientific Conference*, vol. 12, 2825–2830.

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

Pinson, P., C. Chevallier, and G. Kariniotakis (2007): "Trading wind generation from short-term probabilistic forecasts of wind power," in *IEEE Transactions on Power Systems*, vol. 22, 1148–1156.

Pohle, M. (2020): "The Murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation," *arXiv preprint arXiv:2005.01835*.

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

Python Software Foundation (2022): *Python 3.10.4 documentation*.

R Core Team (2021): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rachev, S. T., S. V. Stoyanov, and F. J. Fabozzi (2011): *A Probability Metrics Approach to Financial Risk Measures*, John Wiley & Sons.

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

Ray, E. L., L. C. Brooks, J. Bien, M. Biggerstaff, N. I. Bosse, J. Bracher, E. Y. Cramer, S. Funk, A. Gerding, M. A. Johansson, A. Rumack, Y. Wang, M. Zorn, R. J. Tibshirani, and N. G. Reich (2022): "Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States," *arXiv preprint arXiv:2201.12387*.

Ray, E. L., L. C. Brooks, J. Bien, J. Bracher, A. Gerding, A. Rumack, M. Biggerstaff, M. A. Johansson, R. Tibshirani, and N. G. Reich (2021): "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/.

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

Reich, N. and C. Rivers (2020): "Scientists want to predict COVID-19's long-term trajectory. Here's why they can't." *Washington Post*, published 15 September 2020, https://web.archive.org/web/20201201155324/https://www.washingtonpost.com/outlook/2020/09/15/scientists-want-predict-covid-19s-long-term-trajectory-heres-why-they-cant/.

Reich, N., R. Tibshirani, E. Ray, and R. Rosenfeld (2021): "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/.

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

Reich, N. G., J. Lessler, S. Funk, C. Viboud, A. Vespignani, R. J. Tibshirani, K. Shea, M. Schienle, M. C. Runge, R. Rosenfeld, et al. (2022): "Collaborative hubs: making the most of predictive epidemic modeling," *American Journal of Public Health*, 112, 839–842.

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

Richardson, D. S., H. L. Cloke, and F. Pappenberger (2020): "Evaluation of the consistency of ECMWF ensemble forecasts," *Geophysical Research Letters*, 47, e2020GL087934.

Rizzo, M. L. and G. J. Székely (2016): "Energy distance," *WIREs Computational Statistics*, 8, 27–38.

Robert Koch Institute (2020): "Coronavirus Disease 2019 – Daily Situation Report of the Robert Koch Institute, 18 November 2020," Published at https://web.archive.org/web/20211113091641/https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Nov_2020/2020-11-18-en.pdf?__blob=publicationFile.

——— (2021a): "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://web.archive.org/web/20210331155213/https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/Bericht_VOC_2021-03-31.pdf?__blob=publicationFile.

——— (2021b): "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://web.archive.org/web/20220325173825/https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/Bericht_VOC_2021-02-10.pdf?__blob=publicationFile.

——— (2023): "COVID-19-Hospitalisierungen in Deutschland," Available from: https://github.com/robert-koch-institut/COVID-19-Hospitalisierungen_in_Deutschland doi: 10.5281/zenodo.7527802.

——— (2025): "SURVSTAT@RKI 2.0," https://survstat.rki.de/Content/Instruction/Content.aspx, accessed 16 January 2025.

Saerens, M. (2000): "Building cost functions minimizing to some summary statistics," *IEEE Transactions on Neural Networks*, 11, 1263–1271.

Schefzik, R., J. Flesch, and A. Goncalves (2021): "Fast identification of differential distributions in single-cell RNA-sequencing data with waddR," *Bioinformatics*, 37, 3204–3211.

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

Schwarz, N., H.-J. Hippler, B. Deutsch, and F. Strack (1985): "Response scales: Effects of category range on reported behavior and comparative judgments," *Public Opinion Quarterly*, 49, 388–395.

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

Shafer, G. and V. Vovk (2008): "A Tutorial on Conformal Prediction," *Journal of Machine Learning Research*, 9, 371–421.

Shaked, M. and J. G. Shanthikumar (2007): *Stochastic Orders*, Springer New York.

Sherratt, K., H. Gruson, H. Johnson, R. Niehus, B. Prasse, F. Sandmann, J. Deuschel, D. Wolffram, S. Abbott, A. Ullrich, et al. (2023): "Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations," *eLife*, 12, e81916.

Sillmann, J., V. V. Kharin, X. Zhang, F. W. Zwiers, and D. Bronaugh (2013): "Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate," *Journal of Geophysical Research: Atmospheres*, 118, 1716–1733.

Srivastava, A., T. Xu, and V. K. Prasanna (2020): "Fast and Accurate Forecasting of COVID-19 Deaths Using the SIkJα Model," *arXiv preprint arXiv:2007.05180*.

Taggart, R. (2022): "Evaluation of point forecasts for extreme events using consistent scoring functions," *Quarterly Journal of the Royal Meteorological Society*, 148, 306–320.

Taylor, J. (2021): "Evaluating quantile-bounded and expectile-bounded interval forecasts," *International Journal of Forecasting*, 37, 800–811.

Taylor, J. W. and K. S. Taylor (2021): "Combining probabilistic forecasts of COVID-19 mortality in the United States," *European Journal of Operational Research*.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012): "An overview of CMIP5 and the experiment design," *Bulletin of the American Meteorological Society*, 93, 485 – 498.

Taylor, K. S. and J. W. Taylor (2020): "A Comparison of Aggregation Methods for Probabilistic Forecasts of COVID-19 Mortality in the United States," *arXiv preprint arXiv:2007.11103*.

Thomson, W. (1979): "Eliciting production possibilities from a well-informed manager," *Journal of Economic Theory*, 20, 360–380.

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

Thorarinsdottir, T. L., J. Sillmann, M. Haugen, N. Gissibl, and M. Sandstad (2020): "Evaluation of CMIP5 and CMIP6 simulations of historical surface air temperature extremes using proper evaluation methods," *Environmental Research Letters*, 15, 124041.

Tolksdorf, K., W. Haas, E. Schuler, L. H. Wieler, J. Schilling, O. Hamouda, M. Diercke, and S. Buda (2022a): "ICD-10 based syndromic surveillance enables robust estimation of burden of severe COVID-19 requiring hospitalization and intensive care treatment," *medRxiv*.

——— (2022b): "Syndromic surveillance for severe acute respiratory infections (SARI) enables valid estimation of COVID-19 hospitalization incidence and reveals underreporting of hospitalizations during pandemic peaks of three COVID-19 waves in Germany, 2020-2021," *medRxiv*.

Townsend, J. T. and H. Colonius (2005): "Variability of the MAX and MIN statistic: A Theory of the quantile spread as a function of sample size," *Psychometrika*, 70, 759–772.

UK Department of Health and Social Care (2021): "Reproduction number (R) and growth rate: methodology," Available online at https://web.archive.org/web/20240806124753/https://www.gov.uk/government/publications/reproduction-number-r-and-growth-rate-methodology/reproduction-number-r-and-growth-rate-methodology#section-2.

UK Scientific Pandemic Influenza Group on Modelling (2020): "Medium-term projections and model descriptions," Consensus statement, considered at UK SAGE 66 on 5 November 2020, https://web.archive.org/web/20240523124647/https://www.gov.uk/government/publications/spi-m-o-covid-19-medium-term-projections-explainer-31-october-2020.

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

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

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

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

Wiesel, J. C. (2022): "Measuring association with Wasserstein distances," *Bernoulli*, 28, 2816–2832.

Wilke, C. O. and C. T. Bergstrom (2020): "Predicting an epidemic trajectory is difficult," *Proceedings of the National Academy of Sciences*, 117, 28549–28551.

Winkler, R. (1972): "A decision-theoretic approach to interval estimation," *Journal of the American Statistical Association*, 67, 187–191.

Wolffram, D., S. Abbott, M. An der Heiden, S. Funk, F. Günther, D. Hailer, S. Heyder, T. Hotz, J. van de Kassteele, H. Küchenhoff, et al. (2023): "Collaborative nowcasting of COVID-19 hospitalization incidences in Germany," *PLOS Computational Biology*, 19, e1011394.

Wolter, N., W. Jassat, S. Walaza, R. Welch, H. Moultrie, M. Groome, D. Amoako, J. Everatt, J. Bhiman, C. Scheepers, N. Tebeila, N. Chiwandire, M. du Plessis, N. Govender, A. Ismail, A. Glass, K. Mlisana, W. Stevens, F. Treurnicht, Z. Makatini, N. Hsiao, R. Parboosing, J. Wadula, H. Hussey, M. Davies, A. Boulle, A. von Gottberg, and C. Cohen (2022): "Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study," *The Lancet*, 10323, 437–446.

Wright, F. (1984): "The asymptotic behavior of monotone percentile regression estimates," *Canadian Journal of Statistics*, 12, 229–236.

Wright, M. and A. Ziegler (2017): "ranger: A fast implementation of random forests for high dimensional data in C++ and R," *Journal of Statistical Software*, 77, 1–17.

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

Zou, D., L. Wang, P. Xu, J. Chen, W. Zhang, and Q. Gu (2020): "Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States," *medRxiv*.