GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

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

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

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

Supervisors
Prof. Dr. Thomas Kneib
Nikos Bosse

Submitted on August 1, 2022
to the Faculty of of Business and Economic Sciences at Göttingen University

Contents

List of Figures

Acronyms

cdf cumulative distribution function

CRPS continuous ranked probability score

ECDC European Center for Disease Prevention and Control

WIS weighted interval score

1 Introduction

In recent years evidence in epidemiology - as well other fields - has accumulated that in order to obtain accurate and well calibrated forecasts of targets of interest, such as case numbers of a disease in a given week, it is often advisable to not rely solely on single model outputs, but to rather consider an aggregate of forecasts made by a group of models, widely referred to as ensemble forecasts (give some cites).

The recent COVID-19 epidemic has turned out to be no exception in this regard. In order to obtain an accurate picture of current disease dynamics for decision makers and following similar efforts in the US, in March 2021 the European Center for Disease Prevention and Control (ECDC) has instigated the European Forecast Hub, collating weekly real-time distributional forecasts for short-term incidence COVID-19 cases and deaths from independent modeling teams Hub (2021). It was found that, in general, ensemble forecasts that aggregate all single model outputs into a single common forecast showed more consistent performance than any single model for both case and death incidence forecasts Sherratt and Gruson. (citecitecite-somemore).

However, while evidence for the advantages of employing an ensemble strategy is ubiquitous, a question that remains is which type of model or ensemble structure gives an edge over others, as well as exploring the interaction between ensemble and model structure. Hence, the question is whether we can identify certain individual modeling or ensemble strategies that consistently perform better than others within the European Forecast Hub. Or alternatively - in lieu of succeeding to establish universal dominance statements - whether it is possible to identify certain situations in which some models or ensemble compositions have an edge over others. In this thesis, the focus will lie on three dimensions of this issue:

First, for the goal of eliciting accurate short-term forecasts, it is an ongoing topic of research whether models that are to some extent explicitly epidemiological in nature, that is, seeking to model transmission dynamics in a population, are to be preferred over models that solely rely on past information of the target time series and are thus agnostic to the underlying transmission dynamics (Funk and Abbott). Different diseases and epidemics warrant/require varying approaches in modeling strategy and we thus aim to investigate whether, for the European Forecast Hub, definitive or (more likely) situation-dependent rankings can be established. For example, Bracher et al. (2021a) identified that statistical models, which rely on past time series dynamics for prediction, are consequently slow to respond to changes in trends, raising the question whether compartmental models fare better in this regard. These types of results can then potentially also be leveraged for forecast composition - in fact, Taylor and Taylor (2021) conjecture that during low-incidence periods, compartmental models should perform better than statistical ones and that ensembles consisting purely of these models should consequently exhibit better performance. Lastly, one must not forget that there might

be different requirements and goals for forecasts depending on who is using them and under which circumstances - these differing goals can be captured with the choice of a corresponding scoring rule. To this end, it is conceivable that, for instance, one model type fares better with regard to point accuracy, while another might exhibit better coverage - for example, it could be evident that one model type exhibits more overconfidence in certain situations. We will thus investigate how the preferred model type varies with the choice of scoring rule.

Another central question in past and current analysis of both European and US COVID Hub data has been whether ensembling procedures should discriminate between their potential member models based on merit. That is, if the ensemble should assign higher weights to models that, in terms of the scoring metric of interest, have performed better in the past than other models, as opposed to applying equal weighting for all models. Results on this procedure of performance-based weighting have been somewhat mixed. For predicting number of deaths in the US forecast hub data, Taylor and Taylor (2021) find that, especially for states that exhibit high mortality, performance-based weighting leads to higher prediction accuracy. (Also include paper by Brooks here, which found no advantage.) Conversely, in the European Forecast Hub, Sherratt and Gruson find no significant improvement of weighted methods in comparison with unweighted and, similarly, Bracher et al. (2021a) also find no systematic benefits of the weighted approach for data from Germany and Poland. Taylor and Taylor (2021) identify a possible reason for the shortcomings of weighted approaches: they require comparable records of historical accuracy and are thus challenging to implement in datasets where model availability fluctuates - this is presumably especially an issue for the European Forecast Hub, where it is common to observe large participation gaps for models.

Lacking evidence for an alternative superior ensembling technique, the European Forecast Hub has thus relied on the unweighted mean ensemble, which was then superseded by the unweighted median ensemble due to its higher resistance to outliers (Sherratt and Gruson). Consequently, the question arises whether it is perhaps possible to establish some guidelines - for example pertaining to situational circumstances, model structure or relation between models - in a data setting where it is not feasible/beneficial to rely on hard and fast mathematical rules via weighted ensembles. The hope is that through the investigation of ensemble behavior in response to these issues, we can establish some heuristics for when it is useful to have certain models in an ensemble or, as a softer goal, to simply gain more insight into how ensemble performance varies in response to the aforementioned dimensions.

There are several possible dimensions to investigate with regard to this research question, with some lines of inquiry having come up before in earlier studies. For instance, a worthwhile question to investigate is whether a model can consistently be underperforming (by one or any scoring metric used) in comparison to other individual models and the ensemble, but can nevertheless provide occasional or consistent benefit when included in said ensemble. An easy example would be a model that consistently underpredicts a target: this is of course in

and of itself undesirable, but could provide great benefit to an ensemble that has a tendency to overpredict the target - a similar case was identified in Bosse and Abbott (2021), and it would be interesting to investigate whether such models can be found in the existing pool of the Hub models. Along these lines, there is also the question whether the choice of adding an additional model is somehow dependent on the summary function used to generate the ensemble: in the case of the models investigated by Bosse and Abbott (2021), it turned out to be somehow "safer" to add models in a median than a mean ensemble, presumably as it is more resistant to outliers.

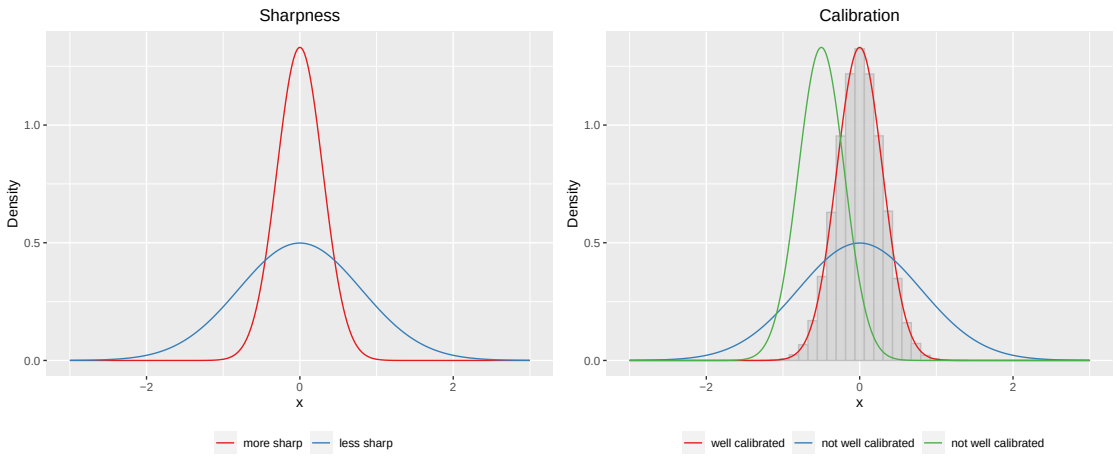
With regard to the aforementioned wide success of ensembles in the forecasting realm, a notion that seeks to explain this success is that averaging over a number of separate models both acts as a mitigator for individual model bias and reduces overall performance variation (do the cites). It is thus conceivable that including models that are too similar and hence somehow make "the same type of mistake" (be it directional, or in relation to over-/ or underconfidence) could, in a sense, hijack/overpower/overtake/skew the ensemble and thus be detrimental to its performance. In turn, this would mean that establishing a notion of "too big" (find more succinct word) model similarity and consequently culling models based on this notion could be beneficial. To this end, we use the Wasserstein 2-metric/Cramer distance, as applied to a discrete set of quantiles and investigate whether excluding single or multiple models that form a sort of "model cluster" improves ensemble performance. Furthermore, we also utilize the existing pool of models to investigate whether subsamples of models with larger overall/average distance were linked to better performance, and also, how performance measures vary as a function of ensemble distance.

Another more basic question we'd like to investigate with regard to ensemble composition is the consistency and variation of forecast performance in relation to the number of its member models. As we believe that including additional models lowers the variation of the ensemble and thus improves its performance, the expectation here is "more is always better" - nevertheless, more knowledge on how exactly ensemble performance relates to ensemble size could be very valuable, especially in situations where resources are limited and it's not immediately clear whether investing into additional models would be rewarding.

As already mentioned, the entire procedure should, to some extent, be regarded more investigatively/inquisitively and with the aim to establish heuristics/soft guidelines rather than hard and fast rules. By nature, the methods described here have a certain ad-hoc character, in a way that having a mathematically formulated rule that "simply" weights by past performance is not. It is possible that no exact guidelines emerge from the analysis, or that emerging results will be very specific to the data at hand and not necessarily generalizable. Nevertheless, we still deem there to be value in this type of analysis, as it can lead to greater understanding of ensemble behavior - performing such an inquisitive deep dive can be regarded as the novel contribution of this thesis.

Finally (note: and if there is time), we want to consider whether tweaking the method of ensemble building might lead to an increase in performance. The aforementioned/currently used methods treat each quantile forecast as separate and thus build a common ensemble forecast by applying some sort of summary function (usually mean or median) to each separate quantile set.[1] However, a potentially sensible/worthwhile/viable alternative approach could be first building a common probability distribution from the set of available forecasts, then taking the quantiles from this aggregate/mixture distribution. To this end, we consider imputing a cumulative distribution function (cdf) for each forecast separately, then taking the ensemble's quantiles from the aggregate/mixture cdf. Put succinctly, we want to investigate whether aggregating the forecasts in cdf rather than quantile space could provide a benefit.
In a nutshell, the goal is of this thesis is to investigate ensemble behavior as it relates to its member models and the current epidemic circumstances, with the hope of potentially deriving some heuristic guidelines for ensemble composition from the findings.

[1] in the case of the median, as identified by Bracher et al. (2021a)), this might lead to not so well formed distributions.



2 Forecasting and Ensembles

2.1 What is forecasting?- the forecasting paradigm

Forecasts vs. scenarios vs. projections.
Sharpness subject to calibration.

2.2 Ensemble Forecasts

They have a long tradition in weather forecasting, where they show consistently improving performance over single models.
An ensemble aggregates models, thereby unifying their respective knowledge/signals into a single forecast.
Ray2020: "Multiple studies of epidemic forecasting have shown that ensemble forecasts, which incorporate multiple model predictions into a combined forecast, consistently perform well and often outperform most if not all individual models (Viboud et al. 2018; Johansson et al. 2019; McGowan et al. 2019; Reich, Brooks, et al. 2019)."

3 Scoring

Suppose that y is the realisation of a random variable under the true data-generating distribution G. The forecasting problem is defined by trying to issue a predictive probability distribution F for the future realisation of this random variable. Further, denote $s(F, G)$ for the expectation of $\mathrm{E}[s(F, y)]$. We then say that scoring rule s is *proper*, if

$$s(G, G) \leq s(F, G).$$

Put into words, this means that the scoring rule is minimized if the true data-generating distribution is issued as the forecast distribution. Likewise, the scoring rule s is *strictly proper*, if

$$s(G, G) < s(F, G).$$

A (strictly) proper scoring rule thus incentivizes the forecaster to issue his or her true belief for the predictive probability distribution.
This notion of the propriety of scoring rules originated with Winkler and Murphy (1968) and its importance in the forecasting world (hmpf) cannot be overstated - if a scoring rule for distributional forecasts is not proper, it could, for instance, incentivize a forecaster to report a more confident estimate than he or she actually believes in Thorarinsdottir 2013.

3.1 Scoring rules

3.1.1 PIT

3.1.2 Coverage

Prediction interval coverage measures the proportion of values that fell into a predictive interval of a given level and thus reflects how well a model was able to characterize uncertainty over time.(Cramer et al.) (from SI, cite something better. Scoringutils paper is also a good reference). It measures probabilistic calibration (Bosse et al., 2022b).

3.1.3 Weighted Interval Score

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

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

where $\mathbb{1}$ is the indicator function, returning 1 if the condition inside the parentheses is fulfilled and 0 otherwise. The three summands each have an intuitive interpretation. The first $(u - l)$ expresses the width of the central prediction interval and thus the sharpness of the predictive distribution F. The second and third summands express under- and over-prediction, respectively. They assign a penalty if the true observed quantity y falls below (above) the lower (upper) endpoint l (u) of the prediction interval. These penalties are furthermore scaled by

the nominal coverage level: a smaller α, which corresponds to a higher nominal coverage rate, induces a higher penalty if y does fall outside one of the endpoints.
Bracher et al. (2021b) extend this score for use on a predictive distribution F that consists of a set of such intervals, each with unique coverage level α. The set of interval scores is gathered and aggregated into the weighted interval score

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

where usually the quantile weights are set to $w_k = \frac{\alpha_k}{2}$, and the median weight to $w_0 = \frac{1}{2}$.
It can be shown that the WIS is an approximation of the continuous ranked probability score (CRPS), a well-known scoring function that measures the distance between the predictive and true distribution

$$CRPS(F,x) = \int_{-\infty}^{\infty}\left(F(y) - \mathbb{1}(y \geq x)\right)^2 dy. \tag{3.3}$$

All in all, the WIS is a parsimonious way to score forecasts that come in the shape of a set of discrete intervals. An important "feature" of the WIS is that it is not standardized: that is, it scales with the data. Scores will increase if the target to be predicted also increases. This makes forecast comparisons a bit difficult, which leads us to the next point.

3.2 Pairwise comparisons

One issue that often arises when aiming to compare different forecasting models is a potentially non-overlapping base of targets the models predicted for, as some scoring rules are not normalized and thus scale with the data. For instance, if models were compared via average WIS, one model might look better than another if it only predicted in periods that saw low incidence or were otherwise comparatively "easy" to forecast. This would thus disincentivize forecasters to predict in periods that they perceive to be more challenging - this is especially undesirable because these periods (e.g. exponential growth, high level of infections) are often of special interest to decision makers (cite something).
One can address this by computing a relative score that is based on employing pairwise comparisons, as developed in Cramer et al.. For a pair of models denoted l and m, first a measure of relative skill is computed

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

where $\bar{s}_l$ and $\bar{s}_m$ denote the average scores the models achieved on the targets both models predicted on - this is usually chosen to be the WIS. For each model, the geometric mean of

these relative scores is then computed as

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

to obtain a relative score of model l with respect to all other available models. It can thus be interpreted as a performance measure of model l with respect to a model with "average" performance. If interest lies in a direct pairwise comparison with a specific model m, one can instead consider the ratio of these relative scores

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

Calculating this ratio for all model pairs that are of interest results in a "pairwise tournament" for all models in the set - this approach is implemented in the **scoringutils** package (Bosse et al., 2022a). For negatively oriented scoring rules, the ratio will be smaller than 1 if model l outperformed model m on their set of shared targets and larger than 1 if it did not. Note that this mode of pairwise comparison still requires the assumption that it is equally hard to perform relatively well to other models at all forecast dates and locations (Cramer et al.).
If one is interested in concisely summarizing the skill of single models rather than performing comparisons between all pairs of models, one can choose a baseline model's B relative score $\theta_{B\cdot}$ as the denominator, which for the WIS results in the measure that is commonly referred to as "relative WIS". Analogously to above, a ratio below 1 corresponds to a model overall outperforming the baseline model, while a score above 1 means that the model did not succeed in clearing baseline performance.

3.3 general

WIS assesses sharpness, while coverage assesses calibration.
Scoring rule is not a meaningless choice: as will be demonstrated in later sections, different scoring rules induce different rankings. It all depends on what the goal is.

4 Data

The data used in this thesis stem from the European forecast hub, which was instigated by the ECDC in 2021 and collates forecasts for Covid-19 cases and deaths from independent modeling teams across Europe (Hub, 2021). Its primary goal is to "provide reliable information about the near-term epidemiology of the COVID-19 pandemic to the research and policy communities and the general public" newer (Sherratt and Gruson). In general, a modeling

hub is a coordinated effort, in which one or more common prediction targets, as well as a common format for prediction, are agreed upon / implemented. This serves the purpose of facilitating model evaluation and development by making model predictions comparable, as well as making predictions suitable for aggregation, that is, for ensemble predictions.
The "hub" format has some precedence both in the realm of climatology as well as in epidemiology, for example in forecasting influenza in the United States Reich and Brooks (2019) as well as dengue fever in ... Johansson et al. (2019). For these seasonal diseases, prediction targets were total number of cases in a season or the height of the peak, while in the case of Covid-19 and the European forecast(ing) hub, the common prediction target are weekly incidence Covid-19 case and death counts in 32 (check) European countries, later also hospitalization rates. Forecasts are issued in a probabilistic manner, namely as a set of 23 quantiles of the predictive distribution, at non-equally-spaced levels between 0.01 and 0.99 (namely $\tau = 0.01, 0.025, 0.05, 0.1, 0.15, ..., 0.85, 0.9, 0.95, 0.975, 0.99$). To be included in the Hub's ensemble and thus in this analysis, models had to provide a full set of 23 quantiles for all four horizons.
Include example plots of individual and ensemble predictions.
Incident deaths are inferred via cumulative. Talk about how the "ensemble is best" paradigm has also held here, with citations to both Eu and Us FCH.
Talk about truth data source, and potential data issues.(incidence is inferred from cumulative).
The Hub also includes a "naive" baseline model, which is the same as the one that is used in the US Covid-19 Forecast Hub (cite). For each forecast date, its forecast for median incidence is equal to the last value for incidence Cases/Deaths that was observed in the most recent week. To obtain the other predictive quantiles, uncertainty around the median is modeled via Monte Carlo approximations of the empirical distribution function that is induced by the first differences observed in the respective time series (Cramer et al.). Predictive quantiles are taken from these samples. Including a baseline model serves the purpose of providing a sort of "minimum" performance that models should be able to clear. Reporting that a model performs better than the baseline thus gives validity to the performance of that model. It's a reference model that all models can be compared against.

4.1 Hub Data

Issued as 23 quantiles. Mostly use WIS, which corresponds to giving slightly larger weight to intervals with large nominal coverage, as compared to CRPS (Bracher et al., 2021b).

5 Introspective stuff

5.1 Model Types

First and foremost, we had to categorize the models in the data in a meaningful manner. We decided on a categorization. To this end, we . Teams that mentioned an explicit compartmental structure of SIR or related (for instance SEIR, SECIR) type we categorized as "mechanistic".
(Bracher et al., 2021b) state that they did not find any "striking patterns" between model types in their analysis, but also acknowledge that this might be due to the relatively short study period they considered. A question that thus naturally arises is, whether given a longer study period, patterns can be found.

6 Ensemble Experiments

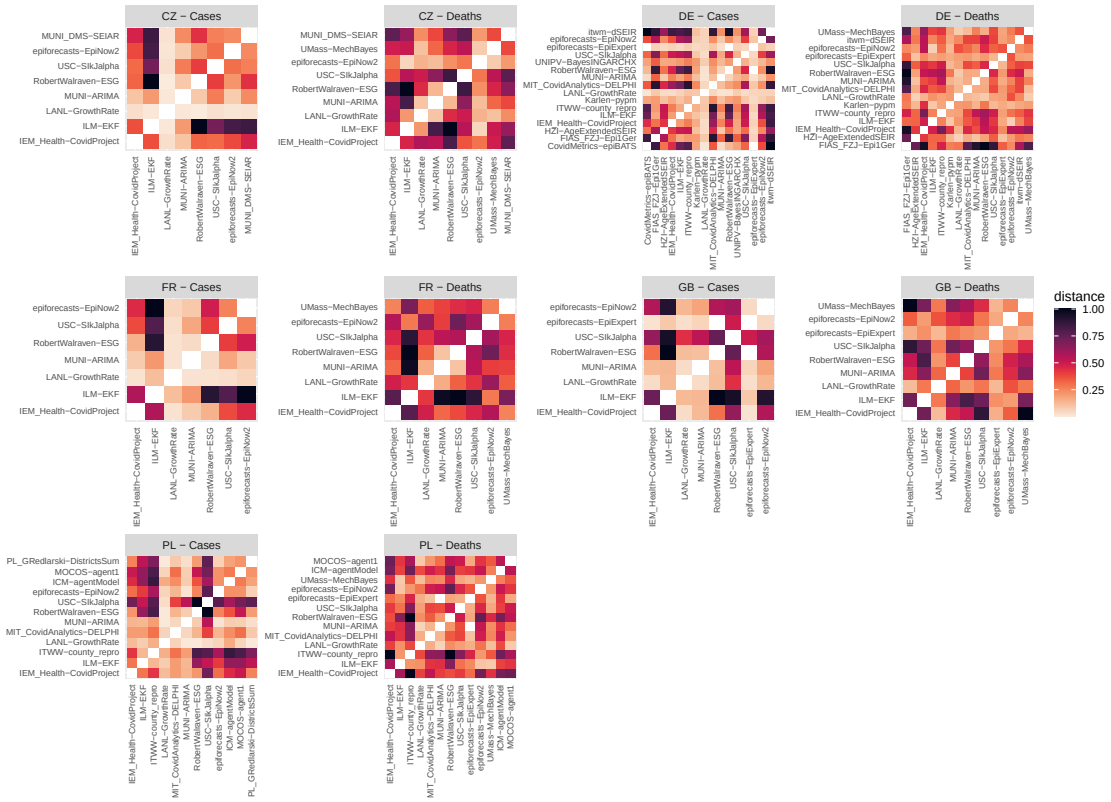
In this section, we

6.1 Model Types

These compartmental models, via a set of differential equations, explicitly model how members of the population transition through the states of being susceptible, (exposed), infected, and recovered/removed Taylor and Taylor (2021). Taylor and Taylor (2021) conjecture that during periods of low incidence, mechanistic models should perform better than statistical ones. This is due to the fact that random statistical fluctuations can still occur, but statistical models might, somehow, latch on to these too eagerly and proceed to forecast exponential growth where there is none.

6.2 Model Similarity

We now turn to the issue of model similarity in ensembles. As expanded upon in Section XX, ensemble models are widely regarded to be successful due to the fact that they counteract/mitigate individual model biases and furthermore reduce variance by aggregating a number of models. Regarding the first point of mitigating bias, it is thus conceivable that ensembling approaches could be less successful if some of the included models are too similar. To illustrate this, recall the , thereby skewing This notion has some mention (find better word) in the literature. For example, in Bosse and Abbott (2021), the authors mention that they purposefully did not submit one of their models for inclusion in the forecast hub's ensemble, as there was concern that it could be too similar to another model they already submitted. However, this decision based on the two models' similarity in modeling setup (shortly explain), rather than



on an actual judgment of how close their predictions were. Nevertheless, they did find that both models improved the ensemble if included (find out if this was actually true). We now want to do a more systematic review of this concept - since we will consider more models across more countries, we hope to get a more accurate picture.

6.3 Weighting based on model types

It is perhaps overeager to assume that one model type could systematically outperform another over the entire study period which, after all, comprises different countries with varying periods of infection dynamics. One could however then conjecture that in specific situations, one modeling philosophy could be better suited than another. For instance, (Taylor and Taylor, 2021) has surmised whether in periods of low infection rates, [ensembles of] compartmental models might be best suited for forecasting, while (Bracher et al., 2021a) have identified that following changes in trends, statistical models perform more poorly, likely because they first have to observe the changing trend in the data from which they extrapolate.

Bibliography

Nikos Bosse and Sam Abbott. Comparing human and model-based forecasts of COVID-19 in Germany and Poland. 2021.

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

Nikos I. Bosse, Hugo Gruson, Anne Cori, Edwin van Leeuwen, Sebastian Funk, and Sam Abbott. Evaluating Forecasts with scoringutils in R. 2022b. doi: 10.48550/ARXIV.2205.07090. URL `https://arxiv.org/abs/2205.07090`. Publisher: arXiv Version Number: 1.

J. Bracher, D. Wolffram, J. Deuschel, K. Görgen, J. L. Ketterer, A. Ullrich, S. Abbott, M. V. Barbarossa, D. Bertsimas, S. Bhatia, M. Bodych, N. I. Bosse, J. P. Burgard, L. Castro, G. Fairchild, J. Fuhrmann, S. Funk, K. Gogolewski, Q. Gu, S. Heyder, T. Hotz, Y. Kheifetz, H. Kirsten, T. Krueger, E. Krymova, M. L. Li, J. H. Meinke, I. J. Michaud, K. Niedzielewski, T. Ożański, F. Rakowski, M. Scholz, S. Soni, A. Srivastava, J. Zieliński, D. Zou, T. Gneiting, M. Schienle, List of Contributors by Team, CovidAnalytics-DELPHI, Michael Lingzhi Li, Dimitris Bertsimas, Hamza Tazi Bouardi, Omar Skali Lami, Saksham Soni, epiforecasts-EpiExpert and epiforecasts-EpiNow2, Sam Abbott, Nikos I. Bosse, Sebastian Funk, FIAS FZJ-Epi1Ger, Maria Vittoria Barbarossa, Jan Fuhrmann, Jan H. Meinke, German and Polish Forecast Hub Coordination Team, Johannes Bracher, Jannik Deuschel, Tilmann Gneiting, Konstantin Görgen, Jakob Ketterer, Melanie Schienle, Alexander Ullrich, Daniel Wolffram, ICM-agentModel, Łukasz Górski, Magdalena Gruziel-Słomka, Artur Kaczorek, Antoni Moszyński, Karol Niedzielewski, Jedrzej Nowosielski, Maciej Radwan, Franciszek Rakowski, Marcin Semeniuk, Jakub Zieliński, Rafał Bartczuk, Jan Kisielewski, Imperial-ensemble2, Sangeeta Bhatia, ITWW-county repro, Przemyslaw Biecek, Viktor Bezborodov, Marcin Bodych, Tyll Krueger, Jan Pablo Burgard, Stefan Heyder, Thomas Hotz, LANL-GrowthRate, Dave A. Osthus, Isaac J. Michaud, Lauren Castro, Geoffrey Fairchild, LeipzigIMISE-SECIR, Yuri Kheifetz, Holger Kirsten, Markus Scholz, MIMUW-StochSEIR, Anna Gambin, Krzysztof Gogolewski, Błażej Miasojedow, Ewa Szczurek, Daniel Rabczenko, Magdalena Rosińska, MOCOS-agent1, Marek Bawiec, Marcin Bodych, Tomasz Ożański, Barbara Pabjan, Ewaryst Rafajłlowicz, Ewa Skubalska-Rafajłowicz, Wojciech Rafajłowicz, Agata Migalska, Ewa

Szczurek, SDSC ISG-TrendModel, Antoine Flahault, Elisa Manetti, Christine Choirat, Benjamin Bejar Haro, Ekaterina Krymova, Gavin Lee, Guillaume Obozinski, Tao Sun, Dorina Thanou, UCLA-SuEIR, Quanquan Gu, Pan Xu, Jinghui Chen, Lingxiao Wang, Difan Zou, Weitong Zhang, USC-SIkJalpha, Ajitesh Srivastava, Viktor K. Prasanna, and Frost Tianjian Xu. A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nature Communications*, 12(1):5173, December 2021a. ISSN 2041-1723. doi: 10.1038/s41467-021-25207-0. URL https://www.nature.com/articles/s41467-021-25207-0.

Johannes Bracher, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17(2):e1008618, February 2021b. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008618. URL https://dx.plos.org/10.1371/journal.pcbi.1008618.

Estee Cramer, Evan Ray, Velma Lopez, and Johannes Bracher. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US.

Sebastian Funk and Sam Abbott. Short-term forecasts to inform the response to the Covid-19 epidemic in the UK.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214506000001437. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

European Covid-19 Forecast Hub. European Covid-19 Forecast Hub. covid19-forecast-hub-europe, 2021. URL Available:https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe.

Michael A. Johansson, Karyn M. Apfeldorf, Scott Dobson, Jason Devita, Anna L. Buczak, Benjamin Baugher, Linda J. Moniz, Thomas Bagley, Steven M. Babin, Erhan Guven, Teresa K. Yamana, Jeffrey Shaman, Terry Moschou, Nick Lothian, Aaron Lane, Grant Osborne, Gao Jiang, Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, Roni Rosenfeld, Justin Lessler, Nicholas G. Reich, Derek A. T. Cummings, Stephen A. Lauer, Sean M. Moore, Hannah E. Clapham, Rachel Lowe, Trevor C. Bailey, Markel García-Díez, Marilia Sá Carvalho, Xavier Rodó, Tridip Sardar, Richard Paul, Evan L. Ray, Krzysztof Sakrejda, Alexandria C. Brown, Xi Meng, Osonde Osoba, Raffaele Vardavas, David Manheim, Melinda Moore, Dhananjai M. Rao, Travis C. Porco, Sarah Ackley, Fengchen Liu, Lee Worden, Matteo Convertino, Yang Liu, Abraham Reddy, Eloy Ortiz, Jorge Rivero, Humberto Brito, Alicia Juarrero, Leah R. Johnson, Robert B. Gramacy, Jeremy M. Cohen, Erin A. Mordecai, Courtney C. Murdock, Jason R. Rohr, Sadie J. Ryan, Anna M. Stewart-Ibarra, Daniel P. Weikel, Antarpreet Jutla, Rakibul Khan, Marissa Poultney, Rita R. Colwell, Brenda Rivera-García, Christopher M. Barker, Jesse E. Bell, Matthew Biggerstaff, David Swerdlow, Luis Mier-y Teran-Romero, Brett M. Forshey, Juli Trtanj, Jason Asher, Matt Clay, Harold S. Margolis, Andrew M. Hebbeler, Dylan George, and Jean-Paul Chretien. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274, November 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1909865116. URL https://pnas.org/doi/full/10.1073/pnas.1909865116.

Nicolas Reich and Logan Brooks. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. 2019.

Katharine Sherratt and Hugo Gruson. (Draft) Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations.

James W. Taylor and Kathryn S. Taylor. Combining probabilistic forecasts of COVID-19 mortality in the United States. *European Journal of Operational Research*, page S0377221721005609, June 2021. ISSN 03772217. doi: 10.1016/j.ejor.2021.06.044. URL https://linkinghub.elsevier.com/retrieve/pii/S0377221721005609.