

# The influence of ensemble size and composition on the performance of combined real-time COVID-19 forecasts

Friederike Becker,

Institute of Statistics, Karlsruhe Institute of Technology

Katharine Sherratt, Nikos Bosse, Sebastian Funk

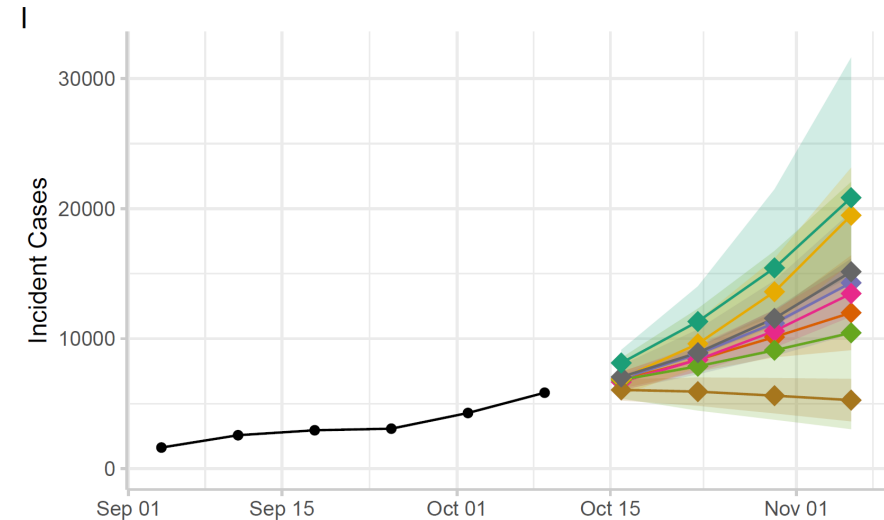
Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine (LSHTM)

Department of Infectious Disease Epidemiology and Dynamics, London School of Hygiene & Tropical Medicine (LSHTM)

SWIM Workshop - December 9, 2025

# European COVID-19 Forecast Hub

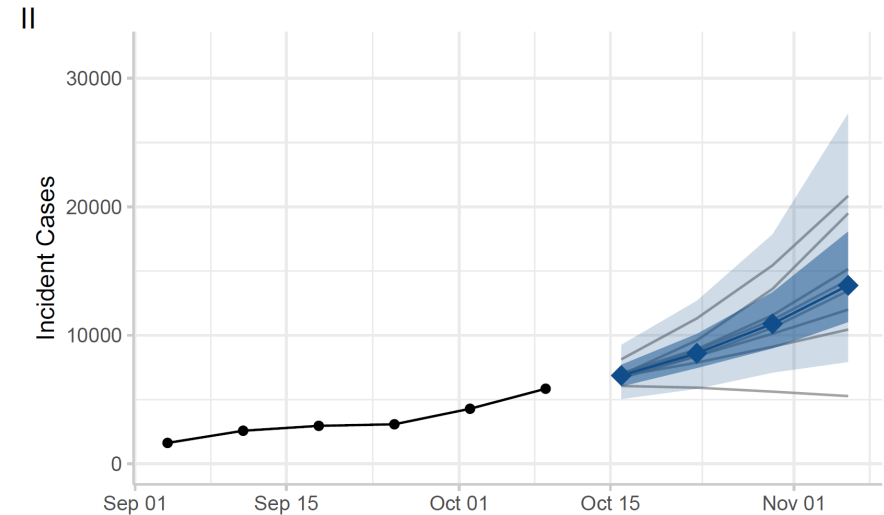
- instigated by the European Centre for Disease Prevention and Control in 2021
- collaborative forecasting effort: collated weekly forecasts from independent modelling teams
- targets: incident COVID-19 Deaths and Cases
- forecast horizons: one to four weeks into the future
- forecast format: probabilistic, via 23 predictive quantiles



Component forecasts for incident cases in the Czech Republic  
forecast origin: Oct 9, 2021

# Models in the Forecast Hub

- forecasts from independent modelling teams - **component forecasts**
  - classified into model types: mechanistic, semi-mechanistic, statistical
  - each component model is usually only available for a subset of locations, targets and time points
- baseline forecast created by the Hub, last value carried forward - **baseline**
- ensemble forecast: quantile-wise equally-weighted median of all component forecasts - **Hub ensemble** or **benchmark**
  - provides more stable predictions than individual models



Hub ensemble forecast for incident cases in the Czech Republic  
forecast origin: Oct 9, 2021

# Research questions

How does ensemble performance relate to **ensemble size** and **ensemble diversity**?

Can ensemble performance be improved by changing **ensemble composition** in favour of well performing component models?

# Research questions

How does ensemble performance relate to **ensemble size** and **ensemble diversity**?

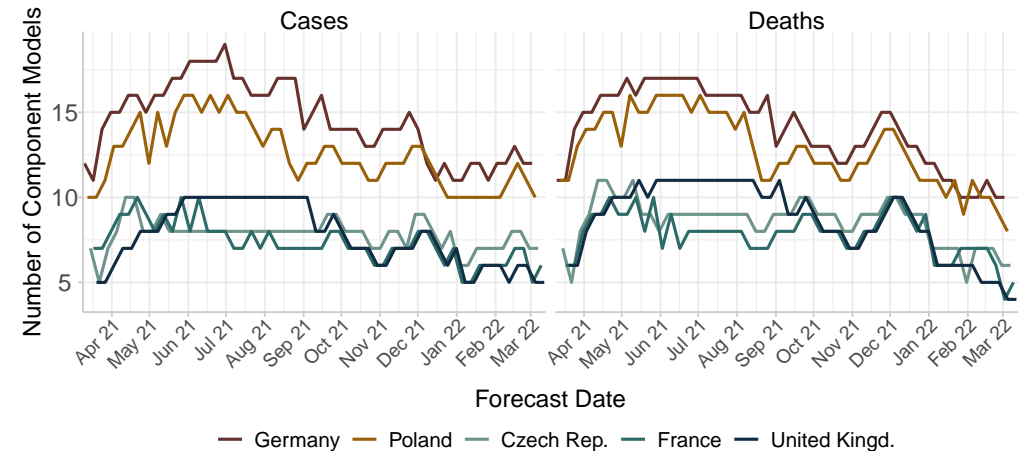
Can ensemble performance be improved by changing **ensemble composition** in favour of well performing component models?

## Previous work

- Fox et al. (2024): *Optimizing disease outbreak forecast ensembles*
- Ray et al (2023): *Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States*
- Amaral et al. (2025): *Post-processing and weighted combination of infectious disease nowcasts*
- Sherratt et al. (2023): *Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations*

# Details on dataset

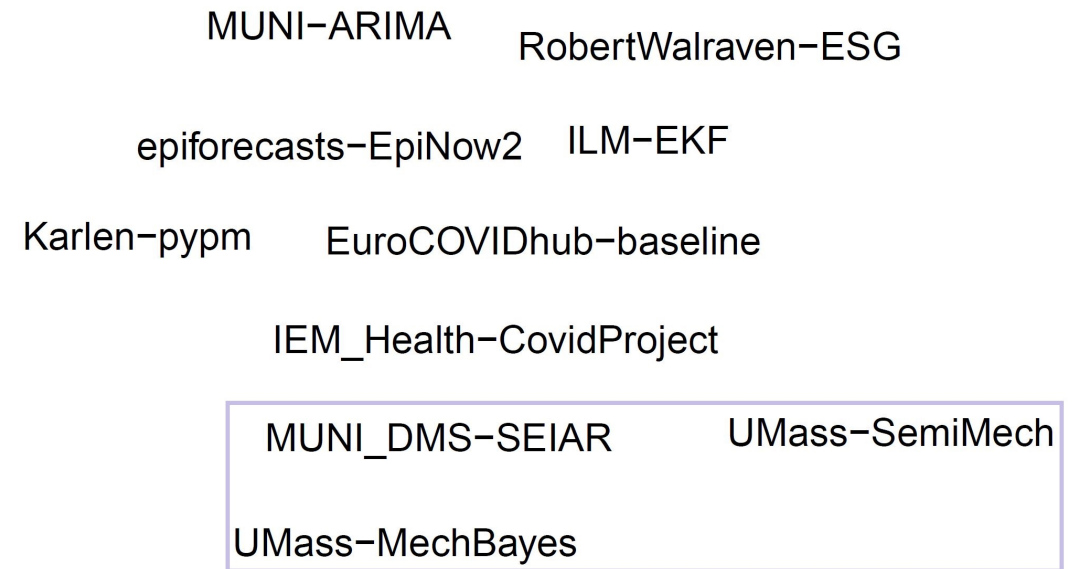
- forecasts issued between Mid-March 2021 and Mid-March 2022 (52 weeks in total)
- five locations with the largest participation rates (Germany, Poland, Czech Republic, France, United Kingdom)
  - participation varies both over time within locations, and between locations
- focus on one and two-week forecast horizons
  - three and four week horizons are used for robustness checks



Number of available models over time

# Ensemble size and diversity - recombination experiment

- by location and target, construct all possible groups of size  $k = 2, \dots, K_j$ , by recombining from available models; then aggregate with the median

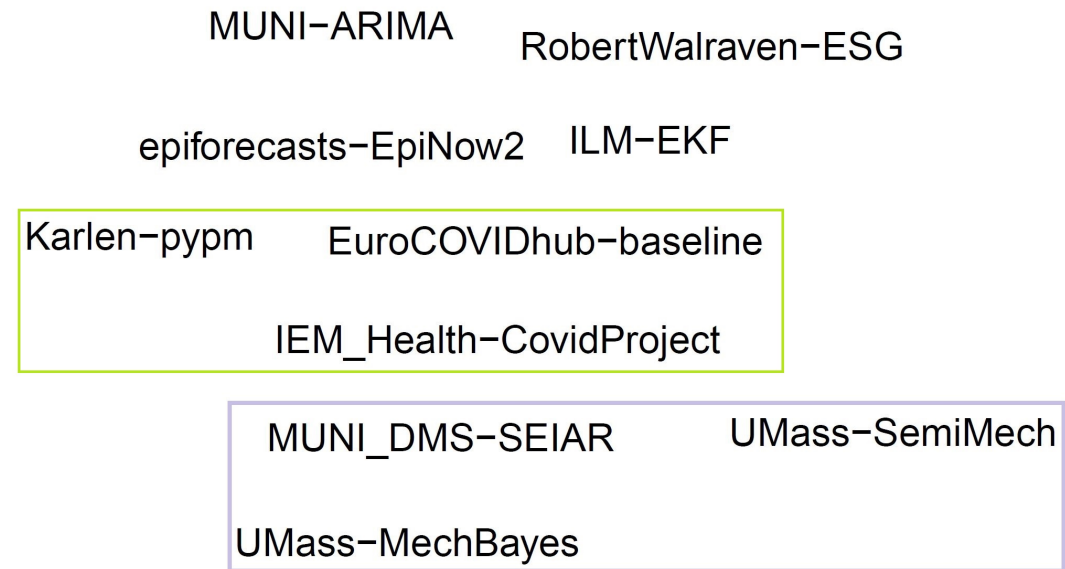


Visualization of recombination process for  $k = 3$

<sup>a</sup>E. Cramer et al. (2022). *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States*, Proc. Natl. Acad. Sci. U.S.A.

# Ensemble size and diversity - recombination experiment

- by location and target, construct all possible groups of size  $k = 2, \dots, K_j$ , by recombining from available models; then aggregate with the median

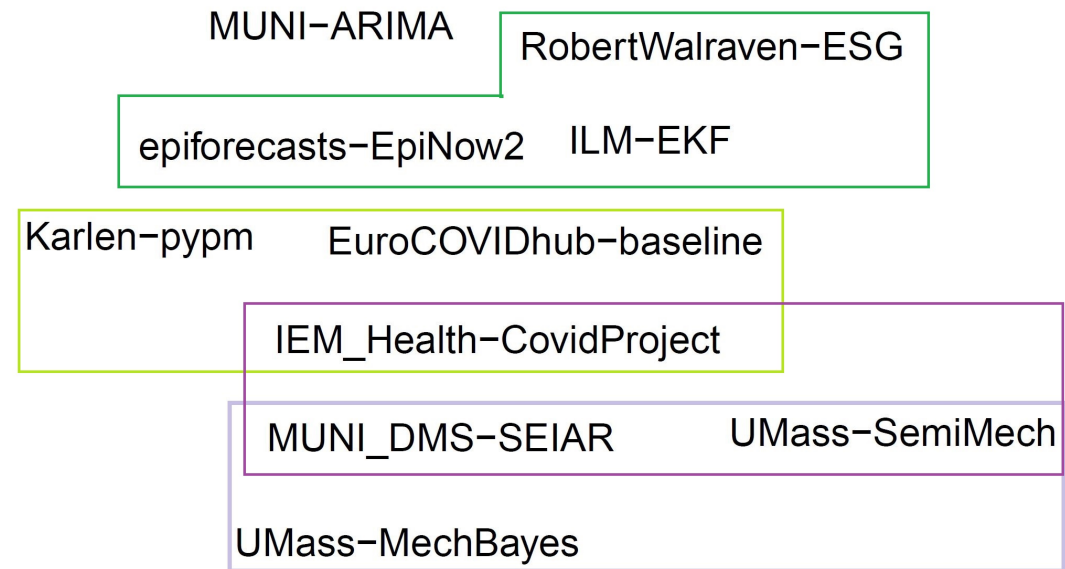


Visualization of recombination process for  $k = 3$

<sup>a</sup>E. Cramer et al. (2022). *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States*, Proc. Natl. Acad. Sci. U.S.A.

# Ensemble size and diversity - recombination experiment

- by location and target, construct all possible groups of size  $k = 2, \dots, K_j$ , by recombining from available models; then aggregate with the median

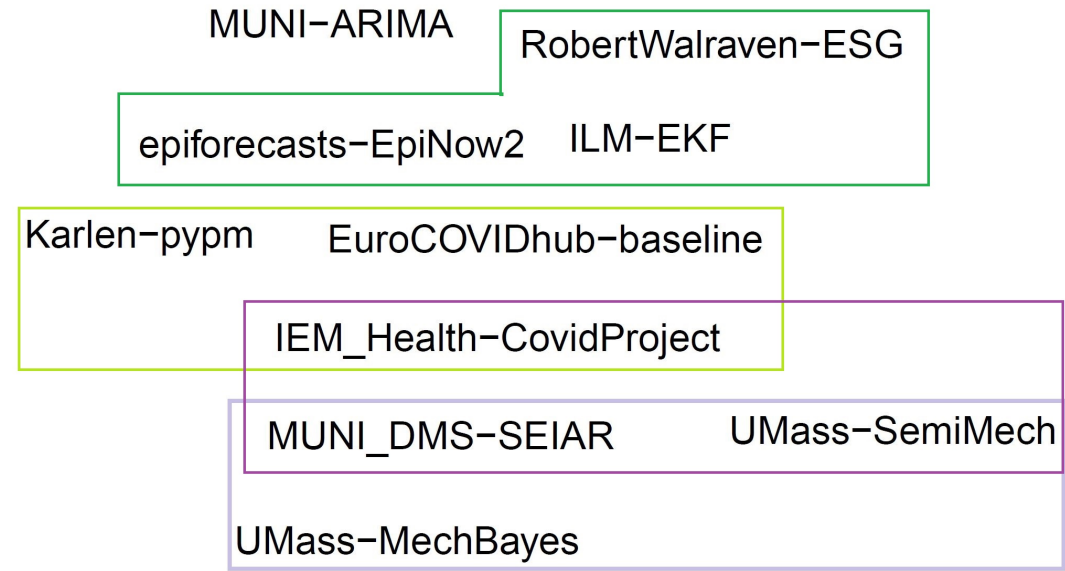


Visualization of recombination process for  $k = 3$

<sup>a</sup>E. Cramer et al. (2022). *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States*, Proc. Natl. Acad. Sci. U.S.A.

# Ensemble size and diversity - recombination experiment

- by location and target, construct all possible groups of size  $k = 2, \dots, K_j$ , by recombining from available models; then aggregate with the median
- score each of the resulting ensembles with the weighted interval score (WIS), relative to the benchmark [link](#)
  - address partial availability by scoring via pairwise comparisons<sup>a</sup> [link](#)

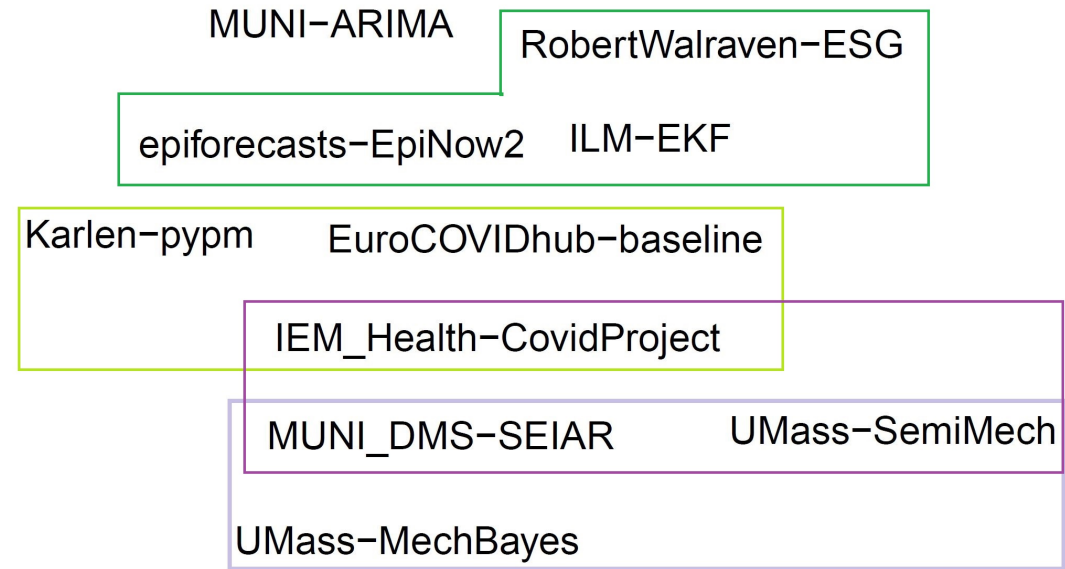


Visualization of recombination process for  $k = 3$

<sup>a</sup>E. Cramer et al. (2022). *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States*, Proc. Natl. Acad. Sci. U.S.A.

# Ensemble size and diversity - recombination experiment

- by location and target, construct all possible groups of size  $k = 2, \dots, K_j$ , by recombining from available models; then aggregate with the median
- score each of the resulting ensembles with the weighted interval score (WIS), relative to the benchmark [link](#)
  - address partial availability by scoring via pairwise comparisons<sup>a</sup> [link](#)
- classify ensembles by degree of component model diversity: heterogeneous and homogeneous

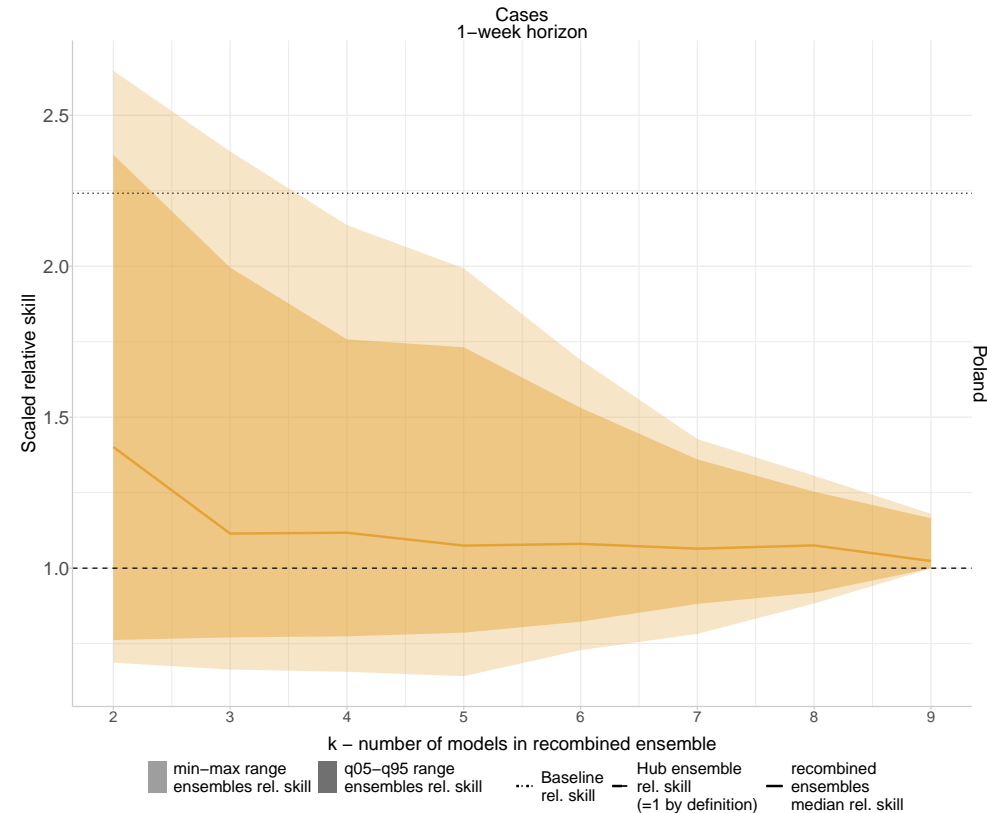


Visualization of recombination process for  $k = 3$

<sup>a</sup>E. Cramer et al. (2022). *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States*, Proc. Natl. Acad. Sci. U.S.A.

# Ensemble size and diversity - results

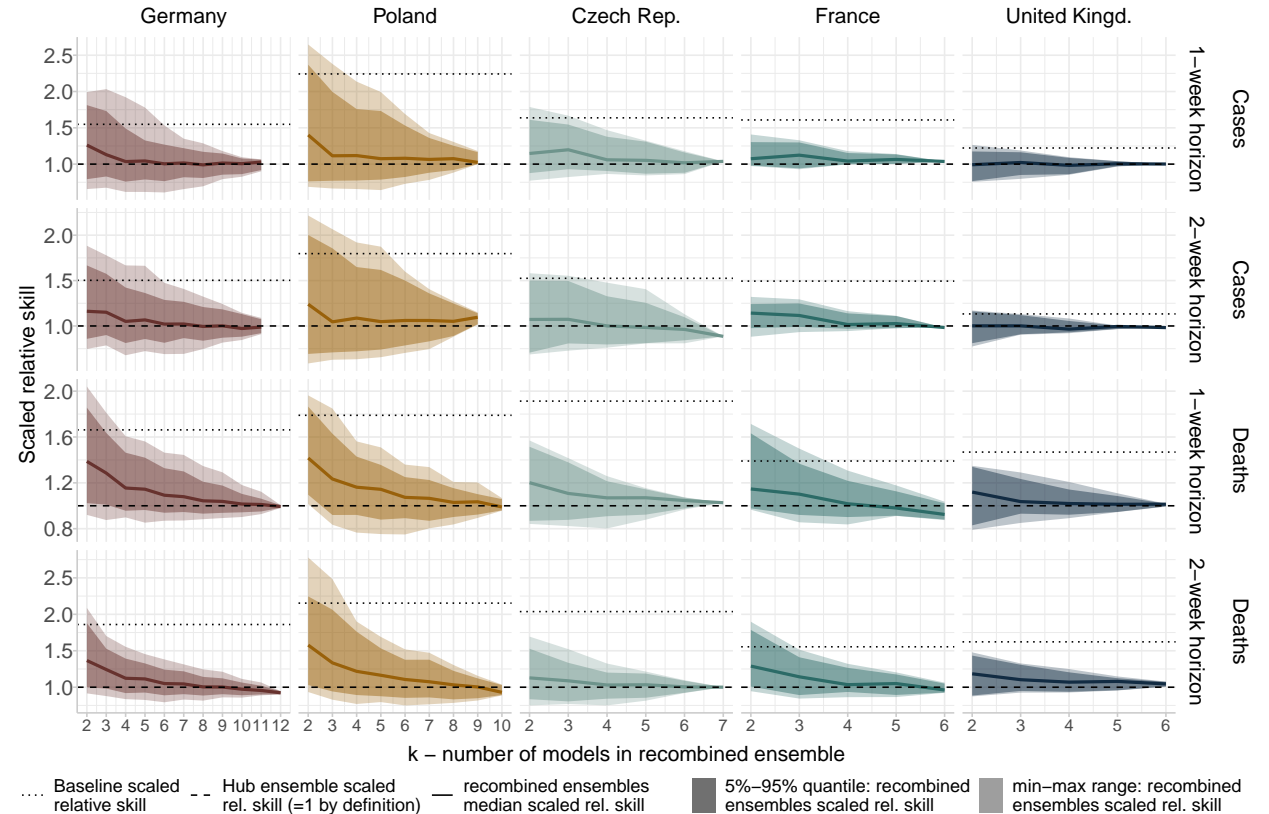
- improvements in median performance often flatten out between three to five component models
- inclusion of more models reduces variability in ensemble performance
- heterogeneous / homogeneous (in terms of model type) ensembles perform similarly



Relative performance of recombined ensembles - Poland Cases

# Ensemble size and diversity - results

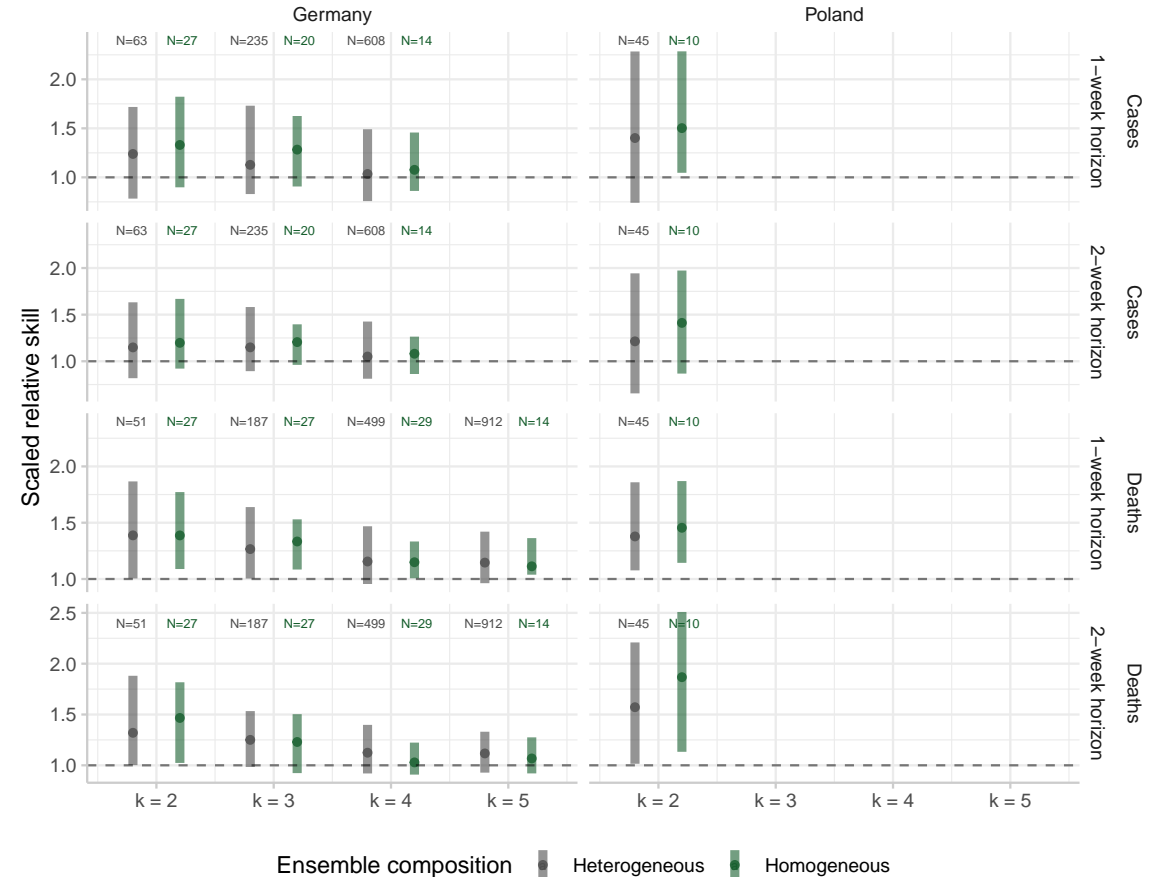
- improvements in median performance often flatten out between three to five component models
- inclusion of more models reduces variability in ensemble performance
- heterogeneous / homogeneous (in terms of model type) ensembles perform similarly



Relative performance of recombined ensembles

# Ensemble size and diversity - results

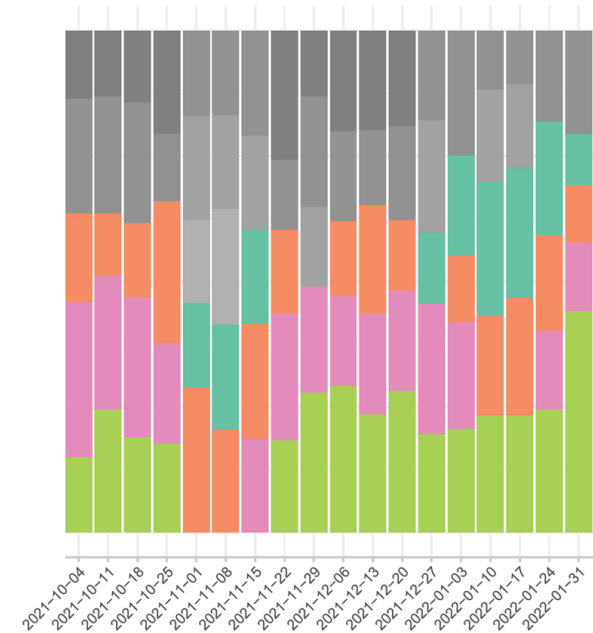
- improvements in median performance often flatten out between three to five component models
- inclusion of more models reduces variability in ensemble performance
- heterogeneous / homogeneous (in terms of model type) ensembles perform similarly



Relative performance of homog./heterog. ensembles

# Ensemble composition - selection process

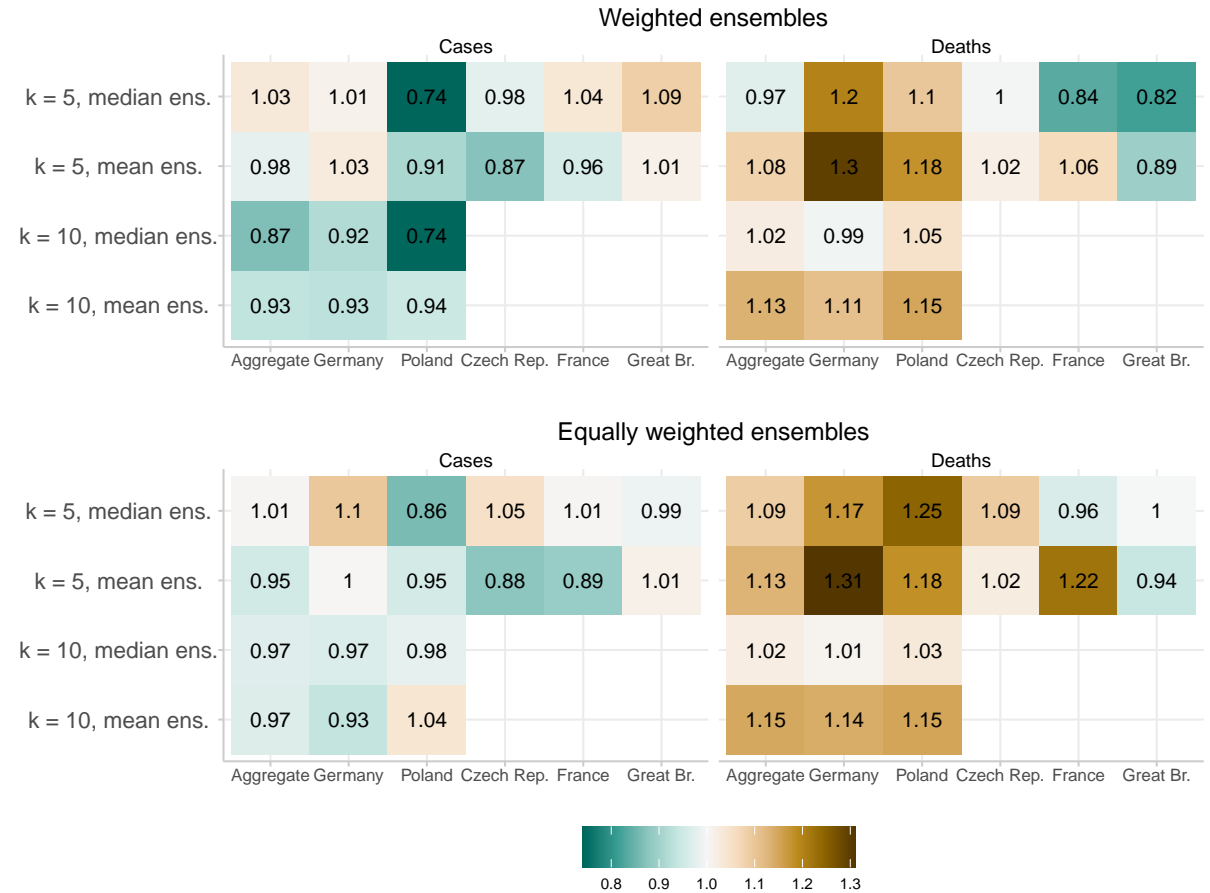
- identify the  $k$  best performing models from the last four weeks
  - $k = 3, 5, 8, 10$
- aggregate best performing models with unweighted mean and median, and weighted mean and median
  - weights are computed via inverse weighted interval scores obtained in the last four weeks → better models receive higher weights
- score relative to the benchmark, using the weighted interval score [link](#)



Example trajectory of weight assignment

# Ensemble composition - results

- improvements in performance are moderate at best, and are not consistent for any method
- especially for Deaths, performance tends to worsen
- variability in chosen models is high
- scores improve and reduce in variability when the ensemble picks more models



Composition ensemble - scores relative to benchmark

# Wrap-up

## Limitations:

- limited size of datasets and number of available models
- ensemble size results may partly reflect an artefact of the recombination process
- there might be more important drivers for model diversity than model type
- more sophisticated composition schemes might still be able to outperform benchmark ensemble

# Wrap-up

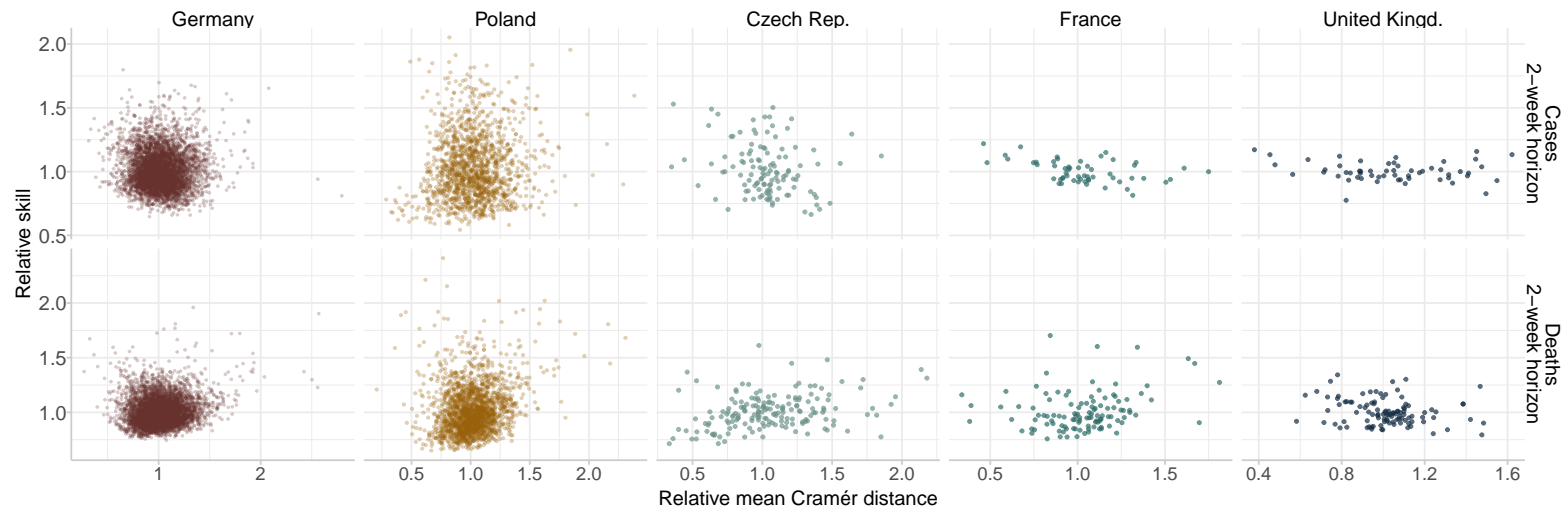
## Limitations:

- limited size of datasets and number of available models
- ensemble size results may partly reflect an artefact of the recombination process
- there might be more important drivers for model diversity than model type
- more sophisticated composition schemes might still be able to outperform benchmark ensemble

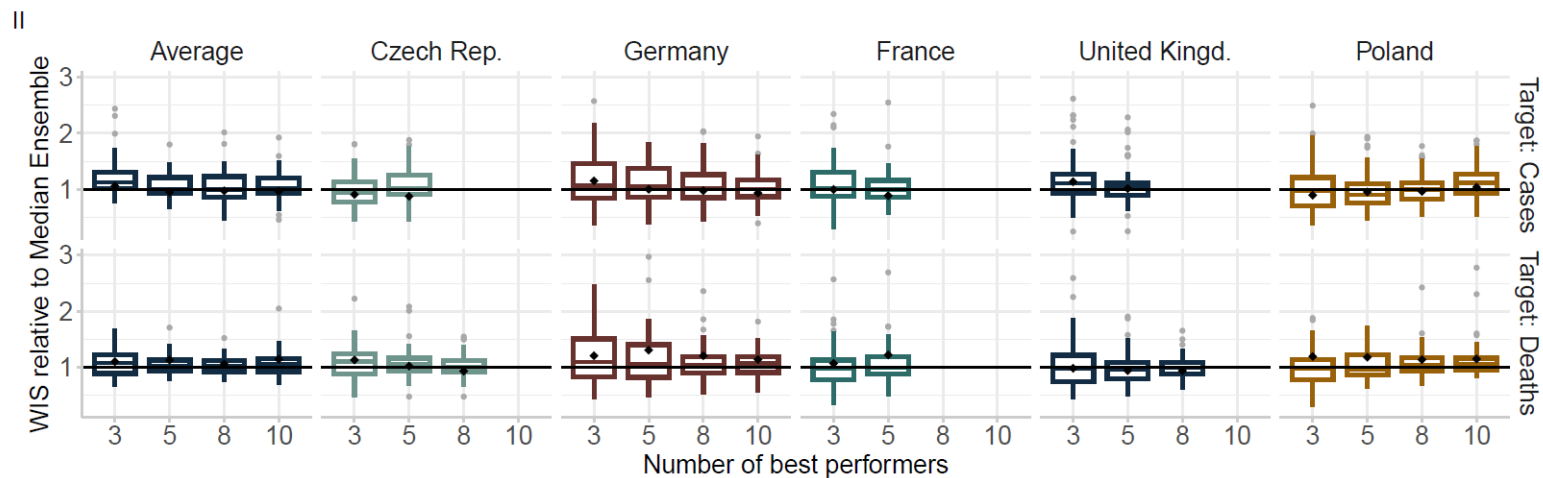
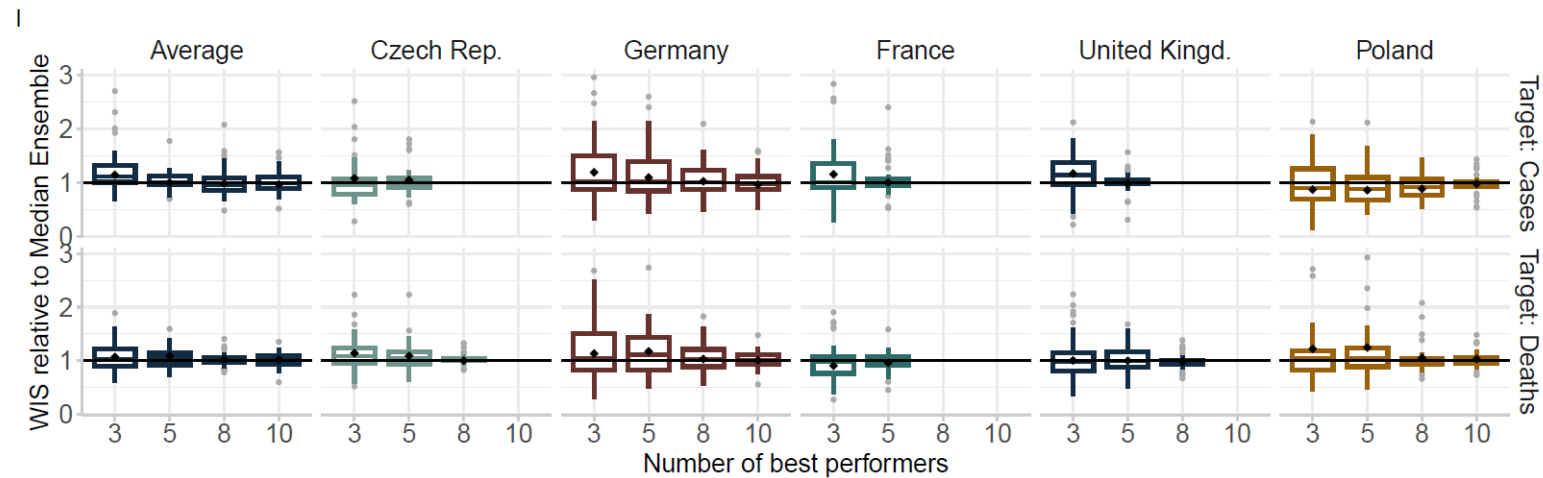
## Key takeaways:

- increasing ensemble size improves performance and reduces variability
  - effect is especially large for up to four or five models
- selecting for better recent performers does not consistently improve ensemble performance
  - another instance of the “forecast combination puzzle”
- practical advice: collate a moderate number of models and aggregate via unweighted median
- results are in line with related prior studies

# Ensemble diversity - numerical measure



# Ensemble composition - boxplots of relative WIS



# Weighted Interval Score

- the Weighted Interval Score (WIS)<sup>1</sup> is a proper scoring rule, meaning that it incentivises forecasters to issue what they believe is the “correct” forecast distribution
- based on the interval score:

$$IS_{\tau}(F, y) = (u - l) + \frac{2}{1 - \tau}(l - y)1(y < l) + \frac{2}{1 - \tau}(y - u)1(y > u), \quad (1)$$

- the WIS is then a weighted sum of all (here: 11) interval scores and the absolute error

» Back

---

<sup>1</sup>Bracher, J. et al. 2021. *Evaluating Epidemic Forecasts in an Interval Format*. PLoS Computational Biology 17 (2)

# Pairwise comparisons

- compute relative scores for all ensemble pairs  $l, m$ ,  
*on overlapping weeks only*

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

- for each ensemble  $l$ , get a mean relative score

$$\theta_{l.} = \text{GM}_{m=1}^M(\theta_{l,m})$$

- scale by *benchmark* to obtain *scaled relative skill*

$$\phi_l = \frac{\theta_{l.}}{\theta_E.}$$

$\phi_l < 1 \cong$  ensemble  $l$  performs better than benchmark

- addresses non-perfect availability by accounting for how difficult it is to beat the benchmark on the targets that ensemble  $l$  addressed

► Back