Results

We evaluated 39 forecast models contributed to the European COVID-19 Forecast Hub, covering 32 countries over 104 weeks from 8 March 2021 to 10 March 2023. Here we report evaluations using log-transformed forecasts. We note this gives less extreme scores compared to using the natural scale (Figure 1; SI Figure 1).

We evaluated forecast performance against the performance of a median ensemble including all qualifying models. We observed the ensemble model generally outperformed individual models. On average across all forecast targets and horizons, only three models outranked the ensemble model (i.e., an average relative WIS <1). Average performance relative to the ensemble varied over time, with a median score of 1.39 and within an interquartile range of 1.07-1.85. Among 104 weeks of forecasts, the median score across all individual models outperformed the score of the ensemble model for only five weeks in autumn 2022.

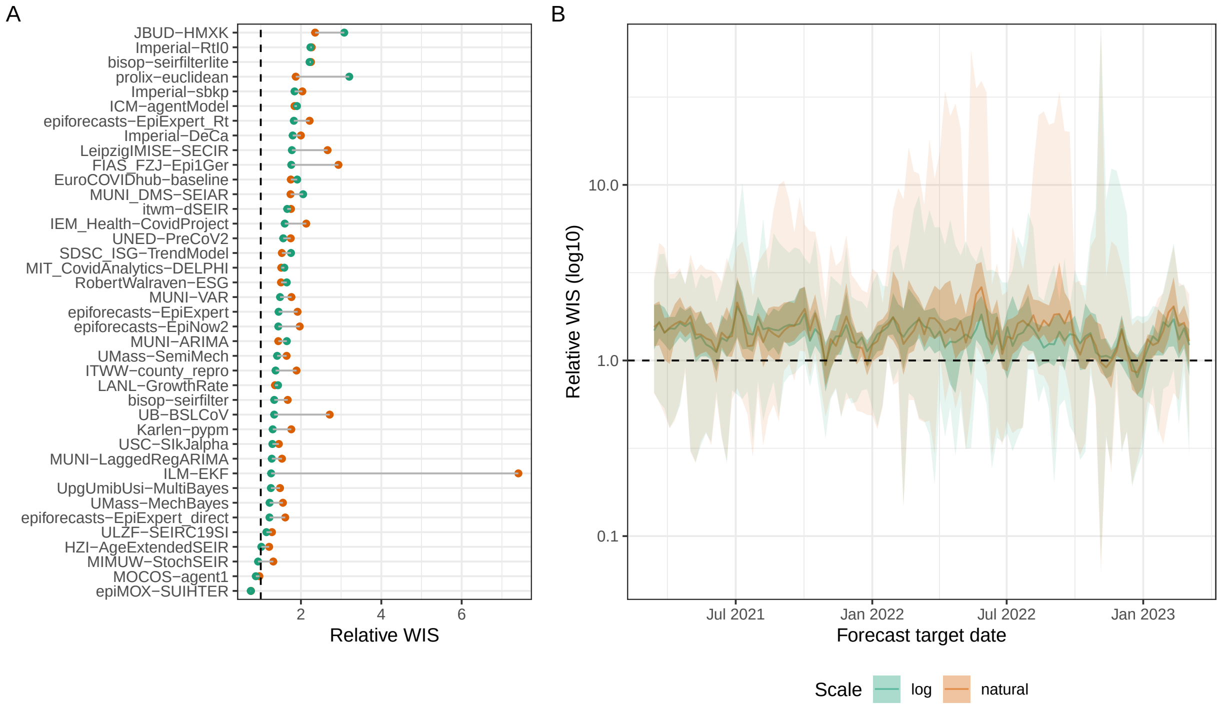

Figure 1: Spread in the individual performance of 39 forecasting models compared to the ensemble model, using scaled relative weighted interval scores (WIS) scored on a log (green) or natural (orange) scale. Relative WIS is the geometric mean of the interval score, scaled relative to the ensemble (dashed line, value 1), with larger scores indicating worse performance relative to the ensemble. The log scale indicates forecasts and observations were transformed with the addition of a constant 1 before scoring. The natural scale indicates raw forecast and observed data were used. (A) Overall scores of each model across all forecast targets (predicting any of 32 locations, up to 4 weeks ahead, for 104 weeks). (B) Interquartile range across 39 models' scores by projection target date, showing median (line), with 50% and 99% quantile interval (shaded ribbons) across models.

Among 39 models, we defined three model structures. We categorised 8 models as statistical, 10 as semi-mechanistic, and 18 as mechanistic, with 3 qualitative ensemble models (table 1). No one model structure consistently outperformed the ensemble (figure 2A). Mechanistic models saw the widest range in performance, with the range of scores including both outperforming the ensemble and performing up to three times worse. Of the three models that, on average, outperformed the ensemble (figure 1), all were mechanistic and forecast for only one country (Poland or Italy).

Table 1. Description of model structures evaluated from the European COVID-19 Forecast Hub.

Structure and examples	Description	Model count
StatisticalARIMA time seriesLOESS seasonal-trend decomposition	Future data predicted entirely from past data. Data may be associated with the forecast target (such as incident cases or human behaviour), but a mechanism for this is not explicitly specified.	8
Semi-mechanistic SIR model with case convolutionState-space model based on the reproduction number	Uses a combination of mathematical equations to describe epidemiological process, combined with data-driven associations	10
Mechanistic SIR modelAge-stratified SEIR model with compartments for progressing disease severityAgent based models: simulations of individuals within a system of defined relationships	Future data predicted entirely from the continuation of an underlying epidemiological process described in a set of mathematical equations. Data used to calibrate the initial model state to the scale of an epidemic, with fixed parameters for the process itself.	18
Qualitative	Qualitative ensembles: A combination of predictions made by models and qualitative human judgement (crowd forecasting).	3

Over the length of the forecast horizon, the performance of statistical models appeared to worsen with increasing horizons up to four weeks. However, this was not the case for mechanistic and semi-mechanistic models. While the median score remained similar over time, the range of forecast scores around it narrowed, indicating more consistent performance with increasing forecast horizon. We noted some variation from these patterns when scores used forecasts on the natural, rather than the log, scale (SI Figure 1).

We considered models forecasting for only one or multiple countries. We collated 23 multi-country models and 16 single-country models. Single-country models targeted Germany, Poland, Czech Republic, Spain, Italy, and Slovenia. Single-country models more often outperformed the ensemble than multi-country models, with a median (99% quantile interval) score of 1.35 (0.77-2.03) compared to 1.6 (1.22-3.17). Performance did not substantially vary with the forecast horizon, with both average and ranges remaining roughly similar for forecasts of one up to four weeks (figure 2B).

Model scores aggregated by target location matching country affiliation

```
## # A tibble: 2 x 6
##   location_match     n  q0.5 q0.01 q0.99 score_range
##   <lgl>          <int> <dbl> <dbl> <dbl> <chr>
## 1 FALSE            650  1.39  0.69  5.33 1.39 (0.69-5.33)
## 2 TRUE              31  1.55  0.77  3.56 1.55 (0.77-3.56)
```

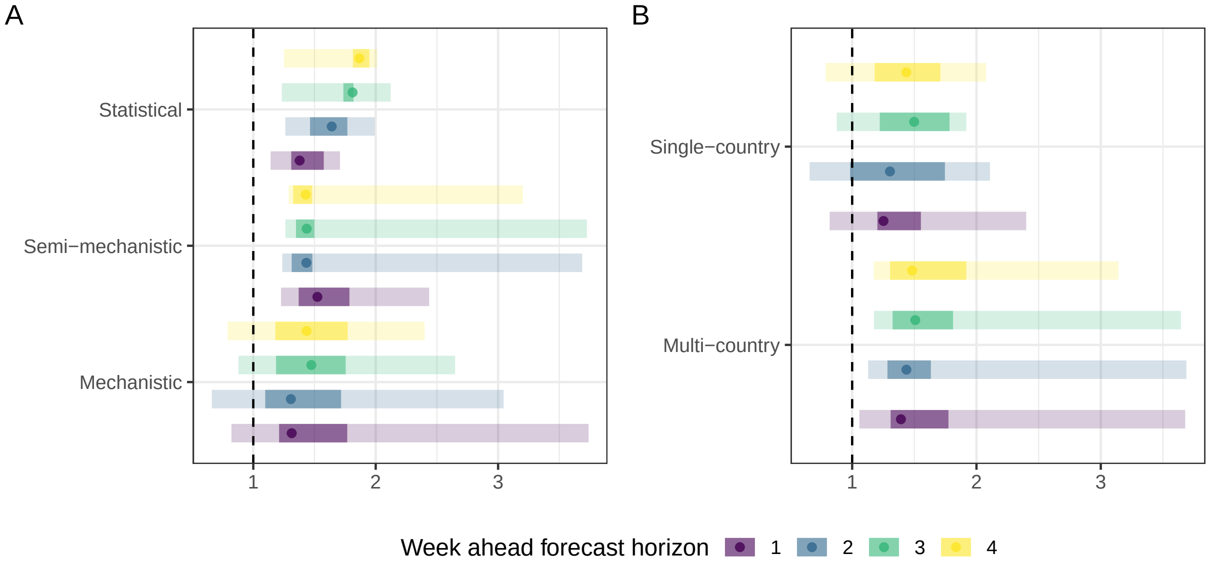

Figure 2: Performance of forecasting models by forecast horizon, by (A) model structure, (B) target specificity, showing the median relative weighted interval score (point), and the 50% and 99% quantile intervals (shaded ribbon), across all available models for each category. Forecasts were log-transformed and scored relative to an ensemble (dashed line, 1).

We fit a generalised additive mixed model to 88141 forecast scores. We found no evidence that forecast performance was affected by either model structure (p=0.48) or the number of countries each model targeted (p=0.45). Performance was more heavily influenced by the observed incidence of deaths from COVID-19, the trend in incidence, and the weeks-ahead horizon of the forecast (each p<0.001; see Supplement). Specifically, a higher observed incidence and an increasing or decreasing trend corresponded to higher interval scores (worsening forecast performance). Similarly, performance declined with longer forecast horizons. This model explains approximately 15.6% of the variability in the relative weighted interval score, suggesting that other factors beyond those included here contribute to forecasting accuracy. We compared models with and without each of the confounding variables included here.

In sensitivity analysis, we found that with one exception, varying the terms included in the GAMM did not influence the substantive result presented here: that neither methodological structure nor number of targets affected forecast performance. However, we noted that most of the variation in forecast performance was absorbed by including each individual model as a random effect (i.e. accounting for the grouping structure of individual forecast scores). By comparison, when this random effect was not included, the impact of methodological structure on forecast performance became substantial (p<0.001), while the relative impact of the number of targets did not change (p=0.02). This suggests that there is greater variation between the performance of each model than there is between their underlying methodological structures.