Results

We evaluated forecasts of incident deaths from COVID-19, collecting 103249 forecasts projected by 39 models contributing to the European COVID-19 Forecast Hub. Forecasts were collected prospectively over 104 weeks from 8 March 2021 to 10 March 2023, and covered one through four week ahead incidence in 32 countries. We report the weighted interval score using log-transformed forecasts.

The number and performance of forecasts varied over time, as forecasting teams joined or left and contributed to varying combinations of forecast targets. We collated between 11 and 33 models in any one week, forecasting for any combination of 128 possible weekly forecast targets. Models widely varied in their volume of contributions: on average each model contributed 2647 forecasts, with the median model contributing 571 forecasts. We observed a range of forecast performance among models (figure 1; Supplement table 1). In general, performance was best in stable periods of little change in incident deaths, while over the length of the forecast horizon, performance appeared to worsen with increasing horizons up to four weeks (table 1).

Table 1: Characteristics of forecasts contributed to the European COVID-19 Forecast Hub, March 2021-2023.

	Models	Forecasts	Median (IQR)
Overall	39 (100%)	103249 (100%)	0.33 (0.16-0.65)
Method			
Mechanistic	18 (46.2%)	34261 (33.2%)	0.31 (0.14-0.61)
Semi-mechanistic	10 (25.6%)	28581 (27.7%)	0.29 (0.16-0.57)
Statistical	8 (20.5%)	38663 (37.4%)	0.4 (0.19-0.74)
Qualitative	3 (7.7%)	1744 (1.7%)	0.23 (0.11-0.49)
Number of country targets			
Single-country	16 (41%)	3933 (3.8%)	0.24 (0.12-0.46)
Multi-country	23 (59%)	99316 (96.2%)	0.33 (0.17-0.66)
Modelling team affiliation			
Affiliated to target country	NA	5111 (5%)	0.26 (0.13-0.51)
Located elsewhere	NA	98138 (95%)	0.33 (0.17-0.66)
Week ahead horizon			
1	39 (100%)	28813 (27.9%)	0.22 (0.11-0.43)
2	34 (87.2%)	25064 (24.3%)	0.3 (0.15-0.56)
3	33 (84.6%)	24820 (24%)	0.39 (0.2-0.72)
4	32 (82.1%)	24552 (23.8%)	0.5 (0.25-0.9)
3-week trend in incidence			
Stable	38 (97.4%)	19377 (18.8%)	0.23 (0.11-0.47)
Increasing	38 (97.4%)	37111 (35.9%)	0.37 (0.18-0.71)
Decreasing	39 (100%)	46437 (45%)	0.35 (0.18-0.67)
NA	12 (30.8%)	324 (0.3%)	0.54 (0.22-1)

We defined four model structures among 39 models. We categorised 8 models as statistical, 10 as semi-mechanistic, and 18 as mechanistic, with 3 qualitative ensemble models. In the volume of forecasts provided, mechanistic, semi-mechanistic, and statistical models each contributed similar numbers of forecasts with approximately one-third each. Descriptively, we observed similar performance in the central tendencies of the

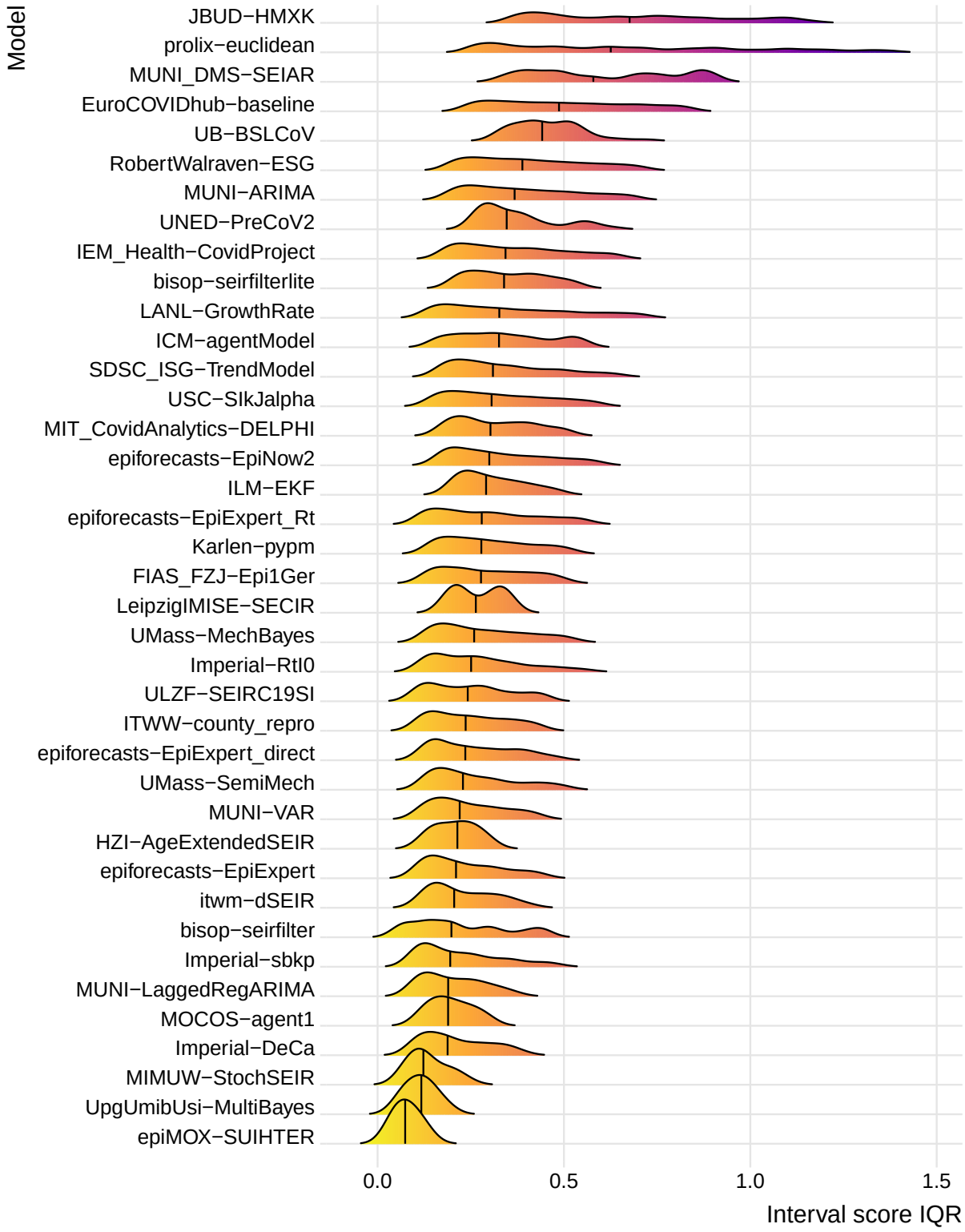

Figure 1: Distribution of forecast scores for 1 to 4 week ahead forecasts across 32 locations over 104 weeks (N=103249). Each distribution shows the interquartile range and median (vertical line) of interval scores across forecasts made by each model, with lower interval score indicating better predictive accuracy.

interval score between mechanistic and semi-mechanistic models, performing relatively better than statistical models. We noted that the four top performing models were all were semi- or mechanistic and forecast for only one country (Poland or Italy), although these models provided far fewer forecasts than others (see Supplement).

We considered models forecasting for one to two, or multiple countries. We collated 16 single-country models and 23 multi-country models. Single-country models targeted Germany (6 models), Poland (5), Czech Republic (2), Spain (2), Italy (2), and Slovenia (1). Two models classified as single-country targeted both Germany and Poland. On average, multi-country models forecast for 24 locations. Models classified as targeting multiple countries could vary from week to week in how many locations they forecast. Only 5 models consistently forecast for the same number of locations throughout the entire study period, with 4 of these forecasting for all 32 available locations.

For each forecast, we associated the target location with the location of the contributing modelling team. Teams were affiliated with 10 of the European countries targeted. 8 teams contributed from outside of Europe (US and Canada), contributing 44.8% of all forecasts. 8 teams were affiliated with the United Kingdom (contributing 23.5% forecasts), with 6 teams each from the Czech Republic (13.2%) and Germany (13.3%). Other team affiliations were to Poland, Spain, Italy, Austria, Switzerland, France, and Slovenia.

We fit a generalised additive mixed model to 101181 forecasts' interval scores. The interval score was highly right-skewed with respect to all explanatory variables (see Supplement). We corrected for this by fitting to the log of the interval score.

We found no clear evidence that any one type of method structure consistently outperformed others (p=0.43). We also found contrasting evidence for whether the location specificity of the model influenced performance. There was a weak indication that targeting multiple countries could be associated with worse forecast performance compared to targeting only one or two countries (p=0.06). At the same time, forecasts for the location in which the modelling team were affiliated appeared to perform worse than those where teams were located elsewhere (p<0.001).

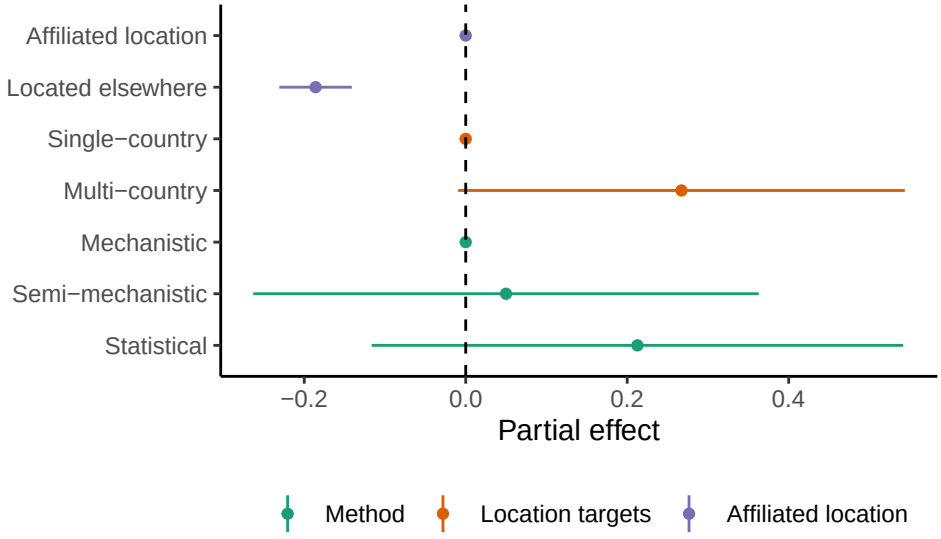

Figure 2: Partial effect size (95% CI) for log-transformed interval score

We compared our results to a null model excluding the three variables of interest. We observed very similar explanatory power (17.6% and 17.5% deviance explained, AIC 275649 and 275714 between explanatory and null models respectively). Among mediating variables, we noted that performance was 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. Considering the model as a whole, this model explained approximately 18% of the variability in the interval score, this might suggest that other factors beyond those included here contribute to forecasting accuracy.