Results

Table 1: Model characteristics contributing to the European COVID-19 Forecast Hub, by method used, number of countries targeted, number of forecasts contributed, and interval scores (median and interquartile range) across contributed forecasts.

Model	Method	Country Targets	Forecasts	Median (IQR)
epiMOX-SUIHTER	Mechanistic	Single-country	134 (0.1%)	0.07 (0.03-0.15)
MIMUW-StochSEIR	Mechanistic	Single-country	76 (0.1%)	0.12 (0.06-0.25)
UpgUmibUsi-MultiBayes	Semi-mechanistic	Single-country	99 (0.1%)	0.12 (0.05-0.19)
MOCOS-agent1	Agent-based	Single-country	386 (0.4%)	0.19 (0.11-0.31)
Imperial-DeCa	Semi-mechanistic	Multi-country	571 (0.6%)	0.19 (0.09-0.4)
Imperial-sbkp	Semi-mechanistic	Multi-country	571 (0.6%)	0.19 (0.09-0.49)
MUNI-LaggedRegARIMA	Statistical	Multi-country	736 (0.7%)	0.19 (0.09-0.38)
bisop-seirfilter	Mechanistic	Single-country	32 (0%)	0.2 (0.07-0.44)
HZI-AgeExtendedSEIR	Mechanistic	Single-country	382 (0.4%)	0.21 (0.12-0.31)
itwm-dSEIR	Mechanistic	Single-country	406 (0.4%)	0.21 (0.11-0.42)
epiforecasts-EpiExpert	Qualitative	Multi-country	948 (0.9%)	0.21 (0.1-0.45)
MUNI-VAR	Statistical	Multi-country	976 (0.9%)	0.22 (0.11-0.45)
UMass-SemiMech	Semi-mechanistic	Multi-country	1904 (1.8%)	0.23 (0.12-0.52)
ULZF-SEIRC19SI	Mechanistic	Single-country	249 (0.2%)	0.24 (0.1-0.45)
ITWW-county_repro	Semi-mechanistic	Single-country	600 (0.6%)	0.24 (0.1-0.44)
epiforecasts-EpiExpert_direct	Qualitative	Multi-country	392 (0.4%)	0.24 (0.12-0.49)
Imperial-RtI0	Semi-mechanistic	Multi-country	571 (0.6%)	0.25 (0.11-0.58)
LeipzigIMISE-SECIR	Mechanistic	Single-country	16 (0%)	0.26 (0.18-0.35)
UMass-MechBayes	Mechanistic	Multi-country	5960 (5.8%)	0.26 (0.12-0.54)
FIAS_FZJ-Epi1Ger	Mechanistic	Single-country	264 (0.3%)	0.28 (0.12-0.52)
Karlen-pypm	Mechanistic	Multi-country	3199 (3.1%)	0.28 (0.13-0.53)
epiforecasts-EpiExpert_Rt	Qualitative	Multi-country	404 (0.4%)	0.28 (0.11-0.58)
ILM-EKF	Semi-mechanistic	Multi-country	12013 (11.6%)	0.29 (0.19-0.5)
MIT_CovidAnalytics-DELPHI	Mechanistic	Single-country	500 (0.5%)	0.3 (0.16-0.52)
epiforecasts-EpiNow2	Semi-mechanistic	Multi-country	7744 (7.5%)	0.3 (0.16-0.61)
USC-SIkJalpha	Mechanistic	Multi-country	12731 (12.3%)	0.31 (0.13-0.61)
SDSC_ISG-TrendModel	Statistical	Multi-country	1755 (1.7%)	0.31 (0.16-0.66)
ICM-agentModel	Agent-based	Single-country	334 (0.3%)	0.33 (0.15-0.57)
LANL-GrowthRate	Semi-mechanistic	Multi-country	3708 (3.6%)	0.33 (0.12-0.73)
IEM_Health-CovidProject	Mechanistic	Multi-country	7720 (7.5%)	0.34 (0.17-0.66)
bisop-seirfilterlite	Mechanistic	Multi-country	336 (0.3%)	0.34 (0.2-0.55)
UNED-PreCoV2	Statistical	Single-country	147 (0.1%)	0.35 (0.26-0.66)
MUNI-ARIMA	Statistical	Multi-country	11369 (11%)	0.37 (0.18-0.69)
RobertWalraven-ESG	Statistical	Multi-country	10488 (10.2%)	0.39 (0.18-0.72)
UB-BSLCoV	Statistical	Single-country	96 (0.1%)	0.44 (0.31-0.72)
EuroCOVIDhub-baseline	Statistical	Multi-country	13096 (12.7%)	0.49 (0.23-0.85)
MUNI_DMS-SEIAR	Mechanistic	Single-country	212 (0.2%)	0.58 (0.33-0.91)
prolix-euclidean	Semi-mechanistic	Multi-country	800 (0.8%)	0.63 (0.24-1.4)

Model	Method	Country Targets	Forecasts	Median (IQR)
JBUD-HMXK	Mechanistic	Multi-country	1324 (1.3%)	0.68 (0.34-1.2)

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.

Among our sample of forecasts, the number 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 both among models and over time (figure 1, supplementary figure 1). As in previous work, we noted that a median ensemble of all forecasts performed consistently well. In general, performance among models 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 2: Characteristics of forecast performance (interval score) 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			
Agent-based	2 (5.1%)	720 (0.7%)	0.23 (0.12-0.44)
Mechanistic	16 (41%)	33541 (32.5%)	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)
Number of country targets			
Qualitative	3 (7.7%)	1744 (1.7%)	0.23 (0.11-0.49)
Single-country	16 (41%)	3933 (3.8%)	0.24 (0.12-0.46)
Week ahead horizon			
Multi-country	23 (59%)	99316 (96.2%)	0.33 (0.17-0.66)
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)
3-week trend in incidence			
4	32 (82.1%)	24552 (23.8%)	0.5 (0.25-0.9)
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. 3 qualitative ensemble models contributed only between March to September 2021. 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 tendency of the interval score between mechanistic and semi-mechanistic models, performing relatively better than statistical models. We noted that the four top performing models

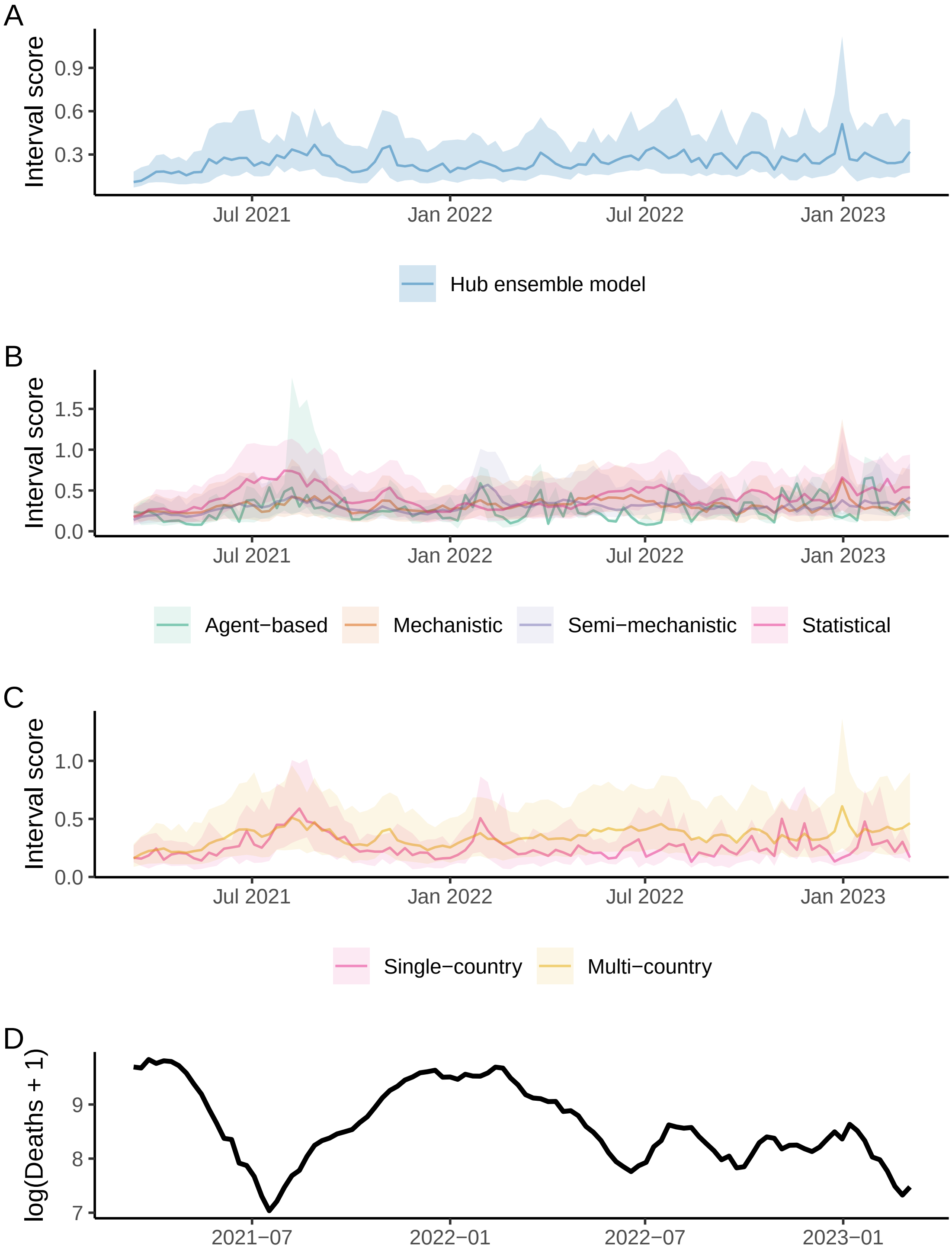

Figure 1: Predictive accuracy of multiple models' forecasts for COVID-19 deaths across Europe. Forecast performance is shown as the median and interquartile range of the weighted interval score, where a lower score indicates better performance. Shown for (A) the Hub ensemble model (the median of all participating forecasts each week); (B) the method used by each model; (C) the number of countries each model targeted (up to 2, or multiple). Forecast performance is summarised across 32 target locations and 1 through 4 week forecast horizons, with varying numbers of forecasters participating over time.

were all semi- or mechanistic and forecast for only one country (Poland or Italy), although these models provided far fewer forecasts than others (table 1, supplementary figure 1). Relative performance among modelling methods also appeared to vary over time (figure 1). For example, statistical models saw a period of poorer performance over summer 2021, coinciding with the introduction of the Delta variant across Europe.

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. Descriptively, multi-country models typically under-performed relative to single-country models to a similar degree over time.

We fit a generalised additive mixed model to 89885 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. We also found no evidence for whether the location specificity of the model influenced performance, comparing models forecasting for three or more countries to those targeting only one or two countries.

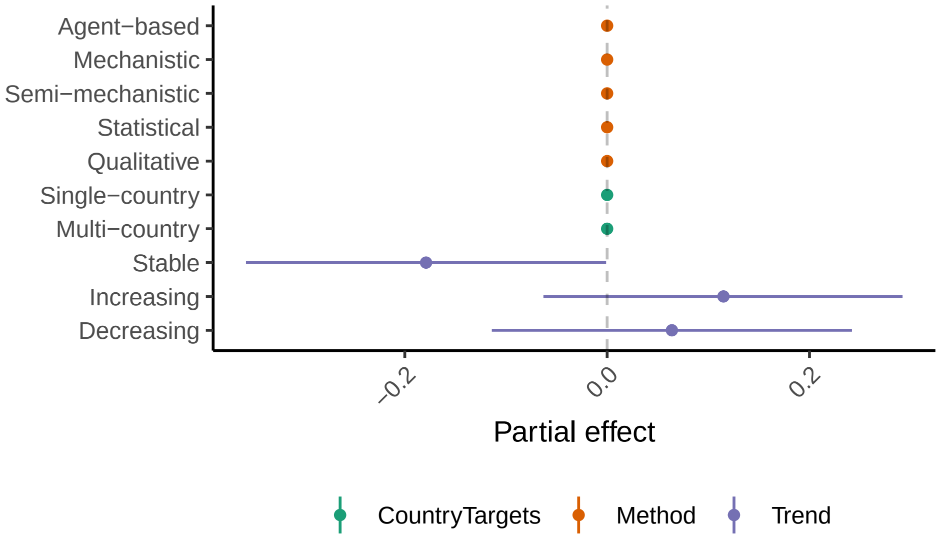

Figure 2: Partial effect size (95% CI) for log-transformed interval score