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	Case forecasts	Death forecasts
AMM-EpiInvert	Statistical	Multi-country	2788 (1.5%)	
CovidMetrics-epiBATS	Statistical	Single-country	343 (0.2%)	
DSMPG-bayes	Semi-mechanistic	Multi-country	760 (0.4%)	
FIAS_FZJ-Epi1Ger	Mechanistic	Single-country	264 (0.1%)	264 (0.1%)
GoeWroc-BaseBayes	Semi-mechanistic	Single-country	12 (0%)	
HZI-AgeExtendedSEIR	Mechanistic	Single-country	382 (0.2%)	382 (0.2%)
ICM-agentModel	Agent-based	Single-country	334 (0.2%)	334 (0.2%)
IEM_Health-CovidProject	Mechanistic	Multi-country	7710 (4.2%)	7708 (4.2%)
ILM-EKF	Semi-mechanistic	Multi-country	11998 (6.6%)	11961 (6.6%)
ITWW-county_repro	Semi-mechanistic	Single-country	650 (0.4%)	600 (0.3%)
Imperial-DeCa	Semi-mechanistic	Multi-country		571 (0.3%)
Imperial-RtI0	Semi-mechanistic	Multi-country		571 (0.3%)
Imperial-sbkp	Semi-mechanistic	Multi-country		571 (0.3%)
JBUD-HMXK	Mechanistic	Multi-country	1324 (0.7%)	1324 (0.7%)
KITmetricslab-bivar_branching	Statistical	Single-country	8 (0%)	
Karlen-pypm	Mechanistic	Multi-country	3208 (1.8%)	3186 (1.8%)
LANL-GrowthRate	Semi-mechanistic	Multi-country	3692 (2%)	3696 (2%)
LeipzigIMISE-SECIR	Mechanistic	Single-country	16 (0%)	16 (0%)
MIMUW-StochSEIR	Mechanistic	Single-country	76 (0%)	76 (0%)
MIT_CovidAnalytics-DELPHI	Mechanistic	Single-country	348 (0.2%)	500 (0.3%)
MOCOS-agent1	Agent-based	Single-country	386 (0.2%)	386 (0.2%)
MUNI-ARIMA	Statistical	Multi-country	10979 (6%)	11314 (6.2%)
MUNI-LaggedRegARIMA	Statistical	Multi-country		736 (0.4%)
MUNI-VAR	Statistical	Multi-country	976 (0.5%)	976 (0.5%)
MUNI_DMS-SEIAR	Mechanistic	Single-country	224 (0.1%)	200 (0.1%)
PL_GRedlarski-DistrictsSum	Mechanistic	Single-country	378 (0.2%)	
RobertWalraven-ESG	Statistical	Multi-country	9190 (5.1%)	10465 (5.8%)
SDSC_ISG-TrendModel	Statistical	Multi-country	1756 (1%)	1744 (1%)
UB-BSLCoV	Statistical	Single-country	96 (0.1%)	96 (0.1%)
UC3M-EpiGraph	Agent-based	Single-country	94 (0.1%)	
ULZF-SEIRC19SI	Mechanistic	Single-country	249 (0.1%)	249 (0.1%)
UMass-MechBayes	Mechanistic	Multi-country		5948 (3.3%)
UMass-SemiMech	Semi-mechanistic	Multi-country	1888 (1%)	1904 (1%)
UNED-PreCoV2	Statistical	Single-country	147 (0.1%)	147 (0.1%)
UNIPV-BayesINGARCHX	Statistical	Multi-country	426 (0.2%)	
USC-SIkJalpha	Mechanistic	Multi-country	12900 (7.1%)	12688 (7%)
UpgUmibUsi-MultiBayes	Semi-mechanistic	Single-country	99 (0.1%)	99 (0.1%)

Model	Method	Country Targets	Case forecasts	Death forecasts
bisop-seirfilter	Mechanistic	Single-country	32 (0%)	32 (0%)
bisop-seirfilterlite	Mechanistic	Multi-country	336 (0.2%)	336 (0.2%)
epiMOX-SUIHTER	Mechanistic	Single-country	134 (0.1%)	134 (0.1%)
epiforecasts-EpiExpert	Qualitative	Multi-country	945 (0.5%)	948 (0.5%)
epiforecasts-EpiExpert_Rt	Qualitative	Multi-country	404 (0.2%)	404 (0.2%)
epiforecasts-EpiExpert_direct	Qualitative	Multi-country	394 (0.2%)	392 (0.2%)
epiforecasts-EpiNow2	Semi-mechanistic	Multi-country	8843 (4.9%)	7721 (4.2%)
epiforecasts-weeklygrowth	Statistical	Multi-country	5974 (3.3%)	
itwm-dSEIR	Mechanistic	Single-country	406 (0.2%)	406 (0.2%)
prolix-euclidean	Semi-mechanistic	Multi-country	800 (0.4%)	800 (0.4%)

We evaluated forecasts of incident deaths from COVID-19, collecting 181854 forecasts projected by 47 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 3869 forecasts, with the median model contributing 764 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.

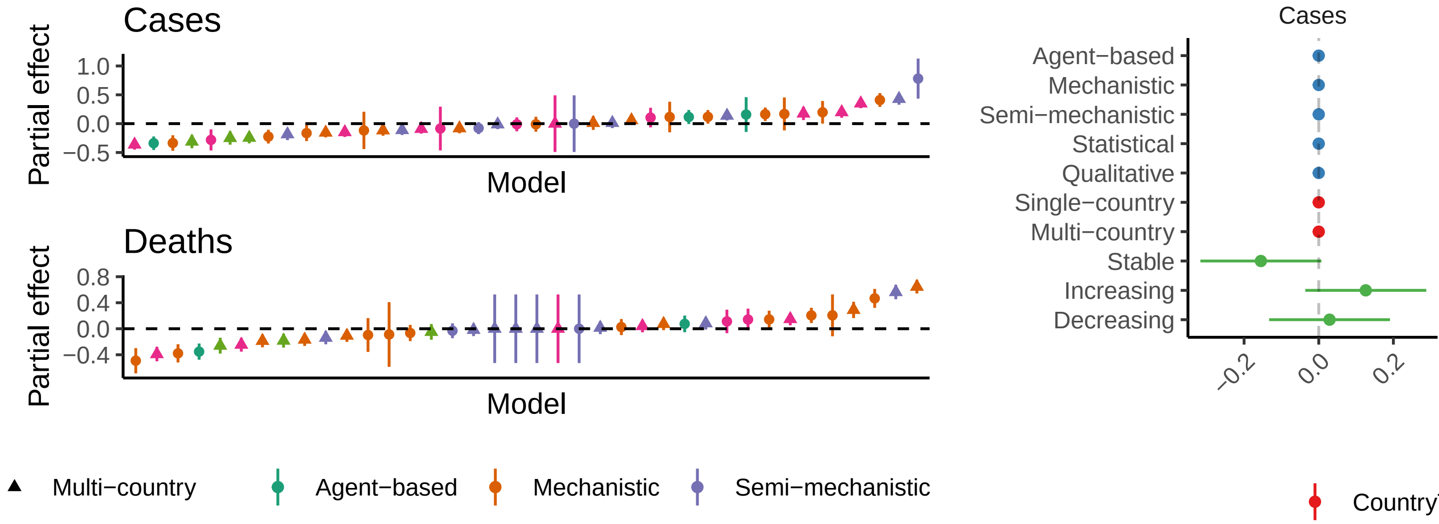
	Cases			Deaths		
Variable	Models	Forecasts	Median WIS (IQR)	Models	Forecasts	Median WIS (IQ
Overall	42 (107.7%)	91969 (100%)	0.28 (0.14-0.63)	38 (97.4%)	89885 (100%)	0.23 (0.12-0.
Method						
Agent-based	3 (7.7%)	814 (0.9%)	0.23 (0.11-0.51)	2 (5.1%)	720 (0.8%)	0.17 (0.09-0.
Mechanistic	16 (41%)	27987 (30.4%)	0.28 (0.12-0.64)	16 (41%)	33449 (37.2%)	0.22 (0.11-0.
Semi-mechanistic	9 (23.1%)	28742 (31.3%)	0.27 (0.14-0.62)	10 (25.6%)	28494 (31.7%)	0.22 (0.13-0.
Statistical	11 (28.2%)	32683 (35.5%)	0.3 (0.15-0.65)	7 (17.9%)	25478 (28.3%)	0.27 (0.14-0.
Qualitative	3 (7.7%)	1743 (1.9%)	0.18 (0.08-0.42)	3 (7.7%)	1744 (1.9%)	0.15 (0.08-0.
Number of country targets						
Single-country	19 (48.7%)	3680 (4%)	0.23 (0.1-0.5)	14 (35.9%)	2821 (3.1%)	0.19 (0.09-0.
Multi-country	23 (59%)	88289 (96%)	0.29 (0.14-0.63)	24 (61.5%)	87064 (96.9%)	0.23 (0.13-0.
Week ahead horizon						
1	42 (107.7%)	24900 (27.1%)	0.14 (0.07-0.28)	38 (97.4%)	25417 (28.3%)	0.17 (0.09-
2	40 (102.6%)	22839 (24.8%)	0.25 (0.13-0.5)	33 (84.6%)	21710 (24.2%)	0.21 (0.12-0.
3	38 (97.4%)	22247 (24.2%)	0.38 (0.2-0.77)	32 (82.1%)	21498 (23.9%)	0.27 (0.15-0.
4	37 (94.9%)	21980 (23.9%)	0.52 (0.26-1.05)	31 (79.5%)	21260 (23.7%)	0.34 (0.19-0.
3-week trend in incidence						
Stable	38 (97.4%)	12684 (13.8%)	0.2 (0.11-0.41)	37 (94.9%)	16793 (18.7%)	0.18 (0.09-0.

Increasing	41 (105.1%)	36674 (39.9%)	0.31 (0.14-0.74)	37 (94.9%)	32402 (36%)	0.25 (0.13-0.
Decreasing	42 (107.7%)	42611 (46.3%)	0.29 (0.15-0.62)	38 (97.4%)	40690 (45.3%)	0.24 (0.14-0.

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 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 23 locations. Models classified as targeting multiple countries could vary from week to week in how many locations they forecast. Only 2 models consistently forecast for the same number of locations throughout the entire study period, with 0 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 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.



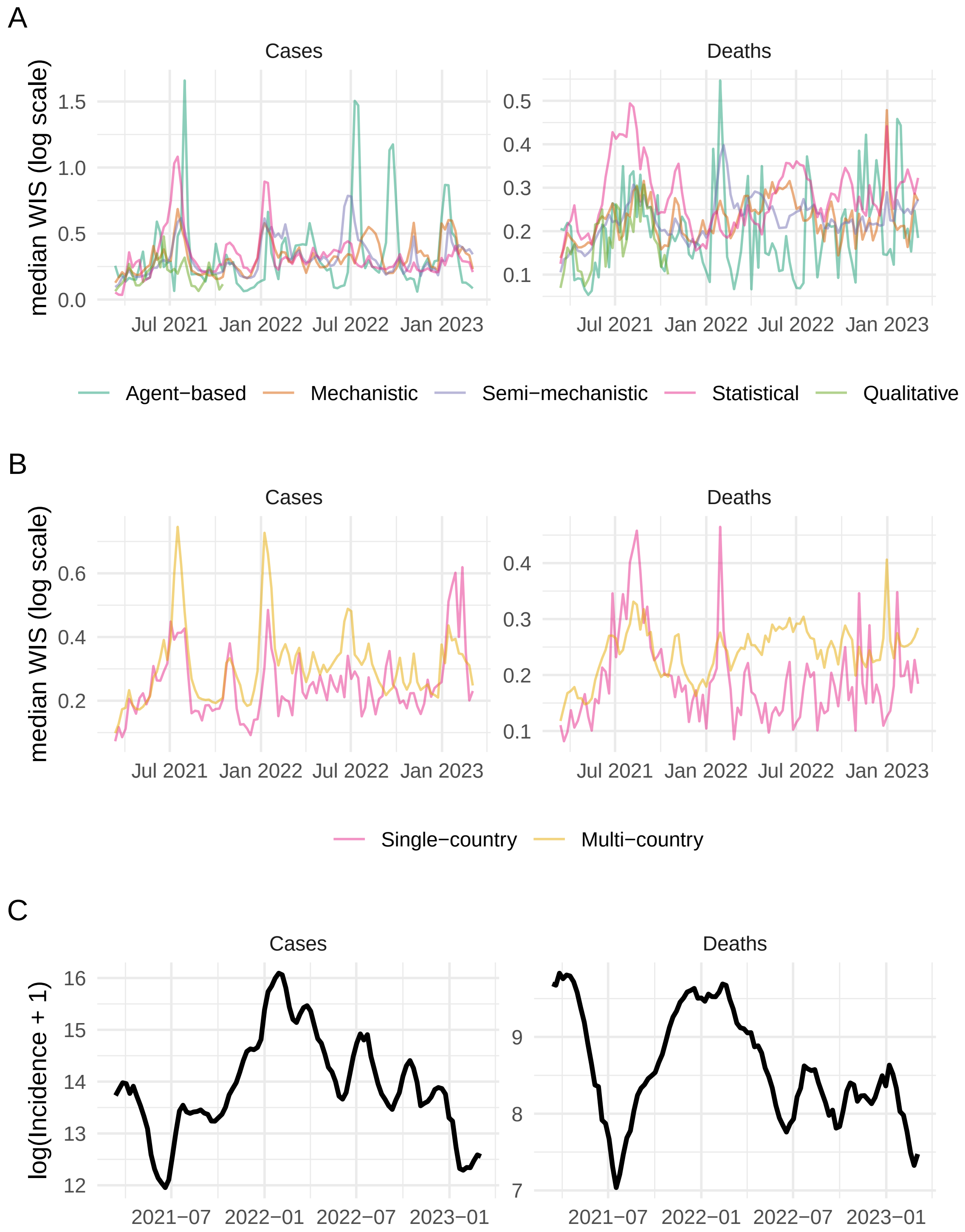


Figure 1: Predictive accuracy of multiple models' forecasts for COVID-19 deaths across Europe. Forecast performance is shown as the median and interquartile range of the weighted interval score, where a lower score indicates better performance. Shown for (A) the Hub ensemble model (the median of all 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.