Results

Model	Method	Country Targets	Forecasts	Median (IQR)	Description
epiMOX-SUIHTER	Mechanistic	Single-country	134 (0.1%)	0.07 (0.03-0.15)	Metadata
MIMUW-StochSEIR	Mechanistic	Single-country	76 (0.1%)	0.12 (0.06-0.25)	Metadata
UpgUmibUsi-MultiBayes	Semi-mechanistic	Single-country	99 (0.1%)	0.12 (0.05-0.19)	Metadata
MOCOS-agent1	Mechanistic	Single-country	386 (0.4%)	0.19 (0.11-0.31)	Metadata
Imperial-DeCa	Semi-mechanistic	Multi-country	571 (0.6%)	0.19 (0.09-0.4)	Metadata
Imperial-sbkp	Semi-mechanistic	Multi-country	571 (0.6%)	0.19 (0.09-0.49)	Metadata
MUNI-LaggedRegARIMA	Statistical	Multi-country	736 (0.7%)	0.19 (0.09-0.38)	Metadata
bisop-seirfilter	Mechanistic	Single-country	32 (0%)	0.2 (0.07-0.44)	Metadata
HZI-AgeExtendedSEIR	Mechanistic	Single-country	382 (0.4%)	0.21 (0.12-0.31)	Metadata
itwm-dSEIR	Mechanistic	Single-country	406 (0.4%)	0.21 (0.11-0.42)	Metadata
epiforecasts-EpiExpert	Qualitative	Multi-country	948 (0.9%)	0.21 (0.1-0.45)	Metadata
MUNI-VAR	Statistical	Multi-country	976 (0.9%)	0.22 (0.11-0.45)	Metadata
UMass-SemiMech	Semi-mechanistic	Multi-country	1904 (1.8%)	0.23 (0.12-0.52)	Metadata
ULZF-SEIRC19SI	Mechanistic	Single-country	249 (0.2%)	0.24 (0.1-0.45)	Metadata
epiforecasts-EpiExpert_direct	Qualitative	Multi-country	392 (0.4%)	0.24 (0.12-0.49)	Metadata
ITWW-county_repro	Semi-mechanistic	Single-country	600 (0.6%)	0.24 (0.1-0.44)	Metadata
Imperial-RtI0	Semi-mechanistic	Multi-country	571 (0.6%)	0.25 (0.11-0.58)	Metadata
LeipzigIMISE-SECIR	Mechanistic	Single-country	16 (0%)	0.26 (0.18-0.35)	Metadata
UMass-MechBayes	Mechanistic	Multi-country	5960 (5.8%)	0.26 (0.12-0.54)	Metadata
FIAS_FZJ-Epi1Ger	Mechanistic	Single-country	264 (0.3%)	0.28 (0.12-0.52)	Metadata
Karlen-pypm	Mechanistic	Multi-country	3199 (3.1%)	0.28 (0.13-0.53)	Metadata
epiforecasts-EpiExpert_Rt	Qualitative	Multi-country	404 (0.4%)	0.28 (0.11-0.58)	Metadata

Model	Method	Country Targets	Forecasts	Median (IQR)	Description
ILM-EKF	Semi-mechanistic	Multi-country	12013 (11.6%)	0.29 (0.19-0.5)	Metadata
MIT_CovidAnalytics-DELPHI	Mechanistic	Single-country	500 (0.5%)	0.3 (0.16-0.52)	Metadata
epiforecasts-EpiNow2	Semi-mechanistic	Multi-country	7744 (7.5%)	0.3 (0.16-0.61)	Metadata
USC-SIkJalpha	Mechanistic	Multi-country	12731 (12.3%)	0.31 (0.13-0.61)	Metadata
SDSC_ISG-TrendModel	Statistical	Multi-country	1755 (1.7%)	0.31 (0.16-0.66)	Metadata
ICM-agentModel	Mechanistic	Single-country	334 (0.3%)	0.33 (0.15-0.57)	Metadata
LANL-GrowthRate	Semi-mechanistic	Multi-country	3708 (3.6%)	0.33 (0.12-0.73)	Metadata
bisop-seirfilterlite	Mechanistic	Multi-country	336 (0.3%)	0.34 (0.2-0.55)	Metadata
IEM_Health-CovidProject	Mechanistic	Multi-country	7720 (7.5%)	0.34 (0.17-0.66)	Metadata
UNED-PreCoV2	Statistical	Single-country	147 (0.1%)	0.35 (0.26-0.66)	Metadata
MUNI-ARIMA	Statistical	Multi-country	11369 (11%)	0.37 (0.18-0.69)	Metadata
RobertWalraven-ESG	Statistical	Multi-country	10488 (10.2%)	0.39 (0.18-0.72)	Metadata
UB-BSLCoV	Statistical	Single-country	96 (0.1%)	0.44 (0.31-0.72)	Metadata
EuroCOVIDhub-baseline	Statistical	Multi-country	13096 (12.7%)	0.49 (0.23-0.85)	Metadata
MUNI_DMS-SEIAR	Mechanistic	Single-country	212 (0.2%)	0.58 (0.33-0.91)	Metadata
prolix-euclidean	Semi-mechanistic	Multi-country	800 (0.8%)	0.63 (0.24-1.4)	Metadata
JBUD-HMXK	Mechanistic	Multi-country	1324 (1.3%)	0.68 (0.34-1.2)	Metadata

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). 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 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			
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)
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 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 (table 1).

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.

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=0.41). 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 (p=0.44).

We compared our results to a null model excluding the three variables of interest. We observed very similar explanatory power (17.5% and 17.5% deviance explained, AIC 275714 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.

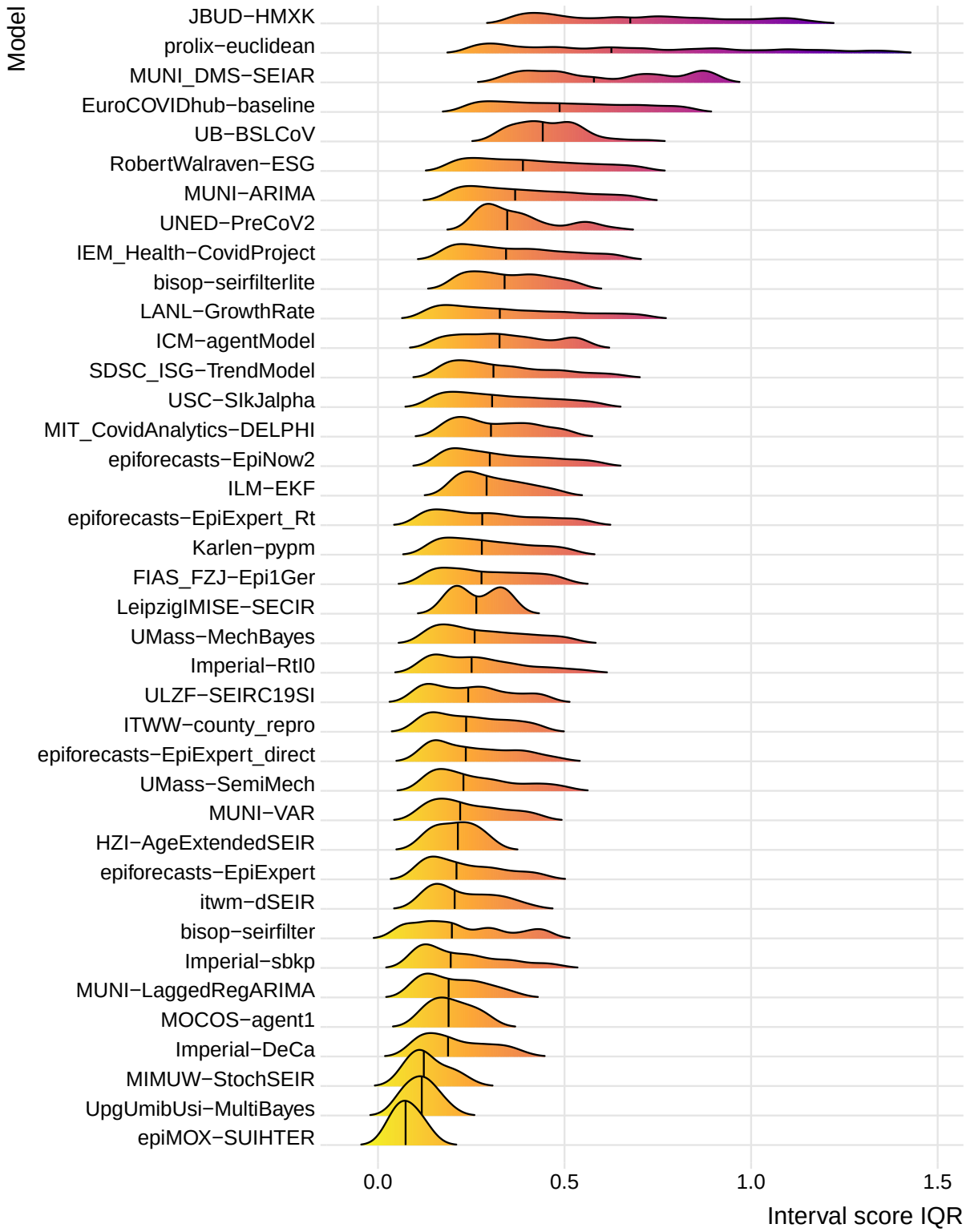

Figure 1: Distribution of forecast scores for 1 to 4 week ahead forecasts across 32 locations over 104 weeks (N=103249). Each distribution shows the interquartile range and median (vertical line) of interval scores across forecasts made by each model, with lower interval score indicating better predictive accuracy. Each model forecast for a different combination of targets, with some models contributing very few forecasts, meaning that forecast scores are not directly comparable.

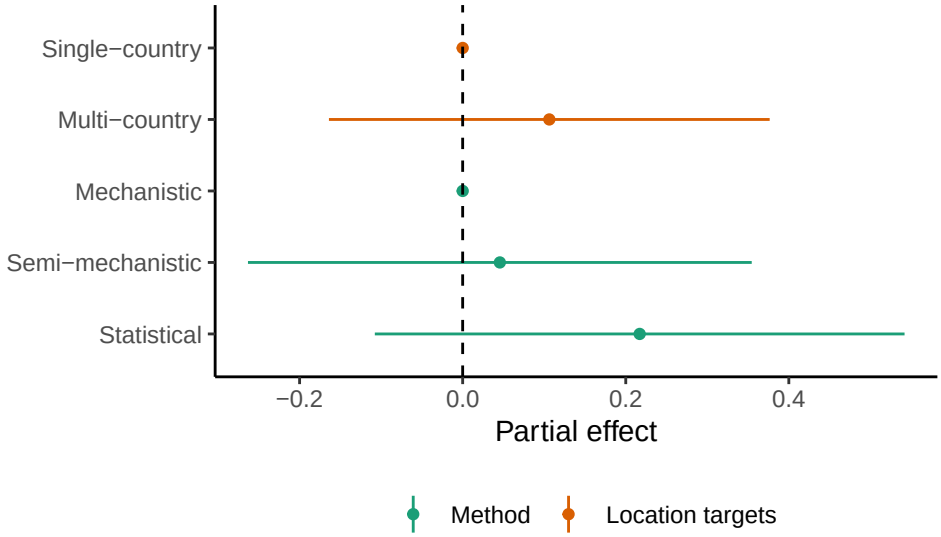

Figure 2: Partial effect size (95% CI) for log-transformed interval score