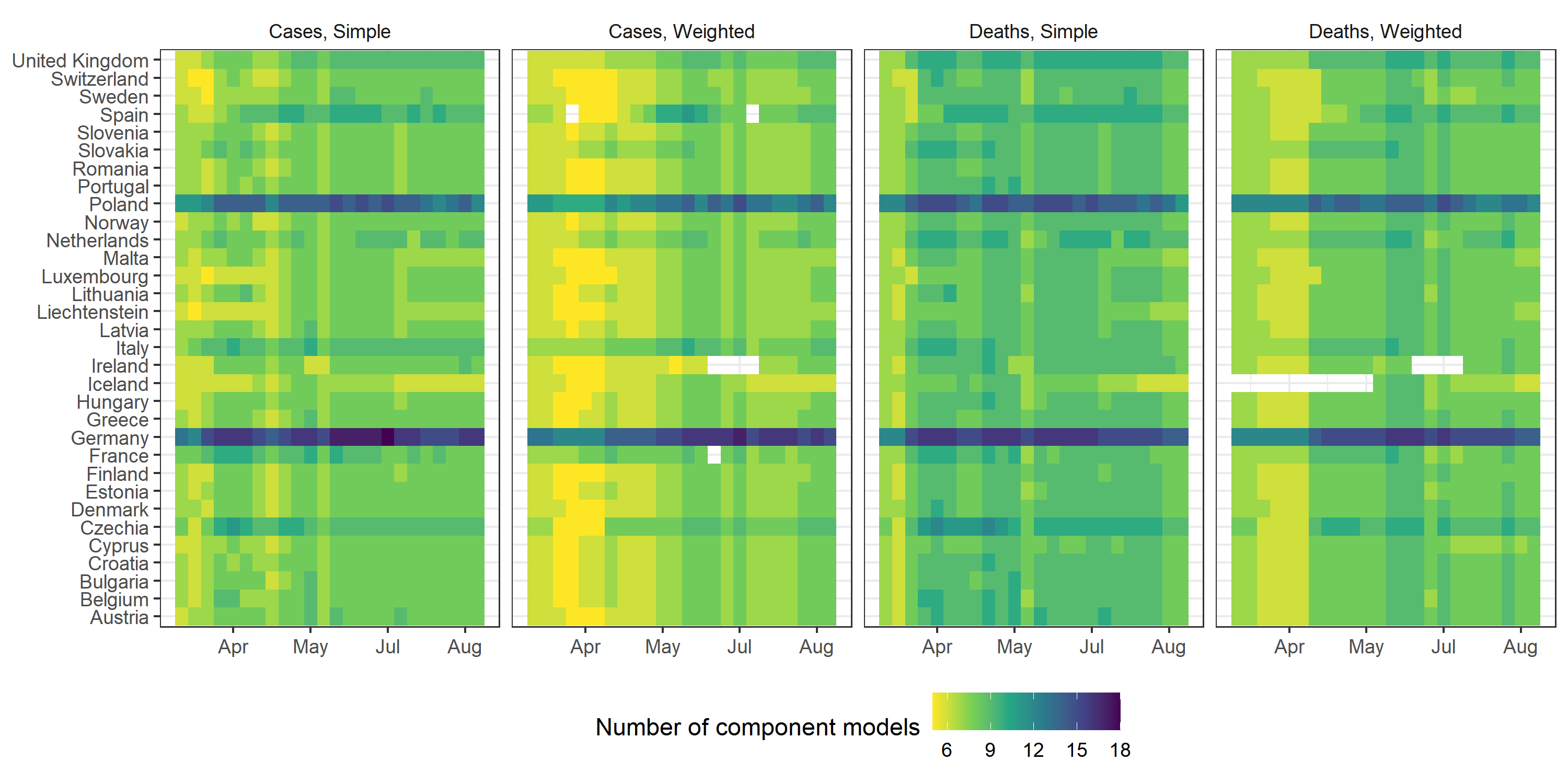
**Results**

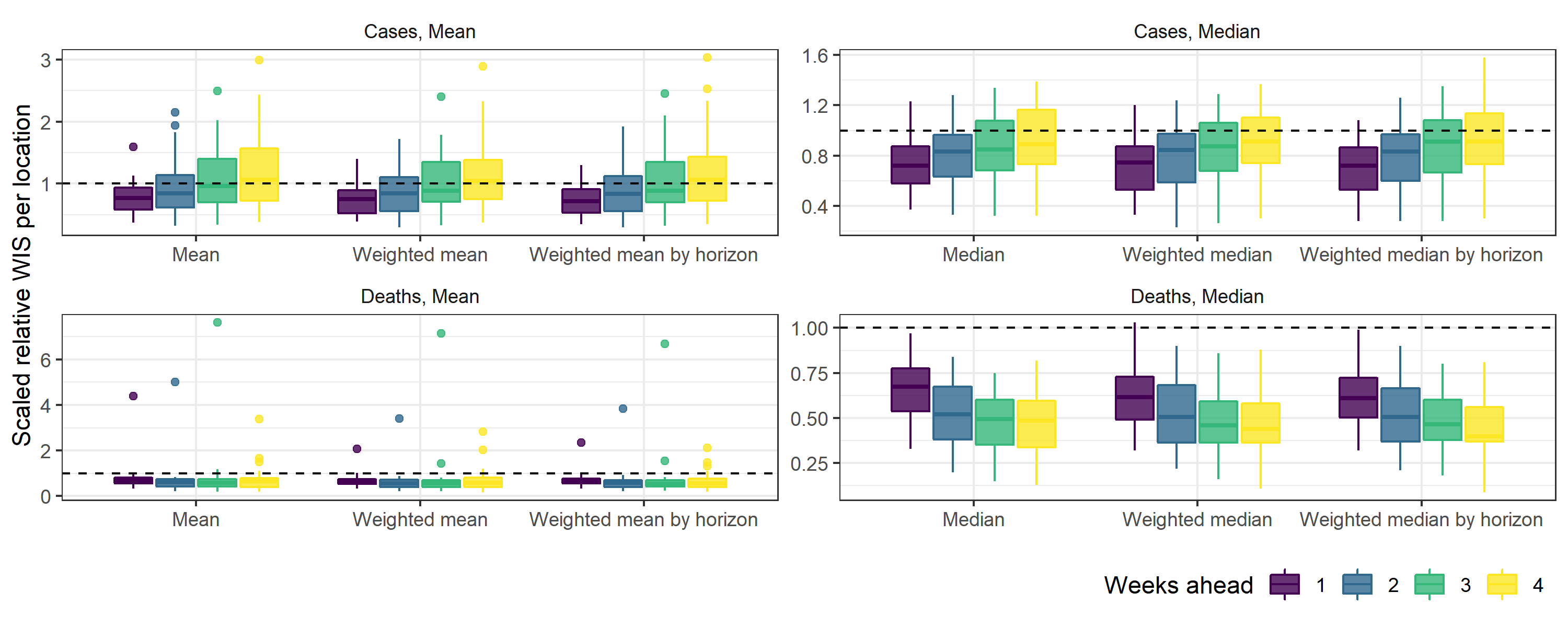
*Component models*



*Figure #1: Number of component forecasts over time for each country, by target count and ensemble method (simple or weighted).*

We created forecasts over the period 2021-03-15 to 2021-08-23. For each week, we created six probabilistic ensemble forecasts of incident weekly case and death counts for 32 countries, forecasting over one through to four weeks (a combined 256 targets). This created 1532 evaluated ensembles, after removing 4 weeks of forecasts in countries reporting data anomalies.

We collected forecasts from a total 29 modelling teams. Ensembles taking the simple averages (mean, median) of all forecasts for each week included between 5 and 18 component models over time. The weighted average ensembles used a stricter set of inclusion criteria, reducing the number of component models to between 5 and 17.



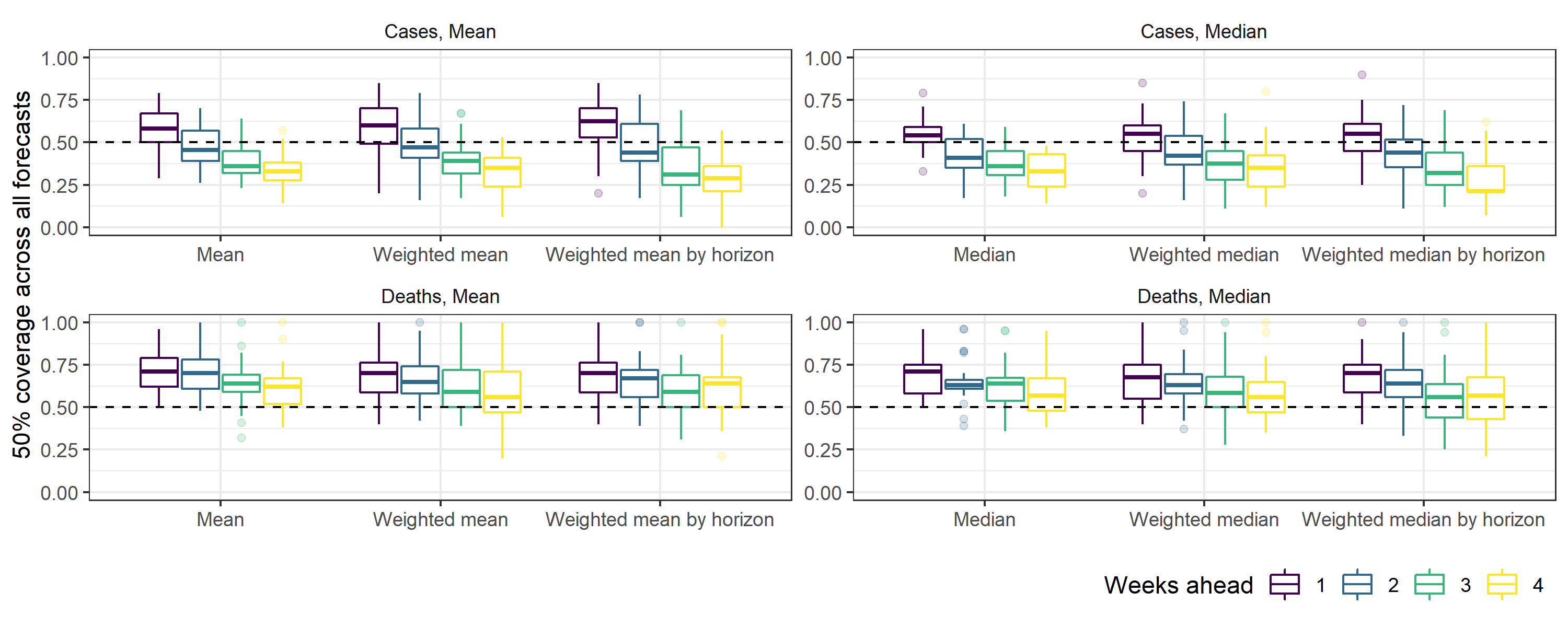
*Figure #2: Relative interval score compared to baseline forecast model (dashed line, 1), by ensemble method (mean or median, with weighted and weighted by horizon methods), target count (cases or deaths), and weeks ahead horizon (1 through 4). Boxplots show distribution of scores across locations, with anomalous ensemble performance (countries with much better or worse forecasts than average) shown as points.*

*Relative performance and uncertainty*

For the majority of forecast targets, ensembles performed better than the baseline. Of the 1532 combinations of the six ensemble methods each forecasting four weeks of case and death counts in 32 countries, 82% performed better than the baseline forecast model.

Ensemble forecasts consistently performed better compared to the baseline model when forecasting incident deaths. With a total 766 targets for incident deaths, 96% ensembles outperformed the baseline model. This was 67% of the same number of case targets.

For all ensemble methods, the skill of ensemble forecasts varied from the near to further into the future, relative to the baseline model. For each model this was always a consistent trend in gradually improving or worsening skill over longer horizons. However the direction of trend varied by the epidemiological target being forecast. Forecasting cases proved more difficult at longer forecast targets than shorter horizons. However, the relative skill of ensembles in forecasting incident deaths improved over longer horizons.

 *Figure #: Coverage, the proportion of observations that fell within the 50% prediction interval for each ensemble, by target count of cases and deaths and horizon. Ideally, a forecast model would achieve 50% of observations within the 50% prediction interval, shown as the dotted line. Values of greater than 0.5 indicate that the forecasts are under-confident (prediction intervals are on average too wide), whereas values smaller than 0.5 indicate that the forecasts are overconfident (prediction intervals tend to be too narrow.)*

When forecasting deaths, ensembles of any model across all horizons typically gave too broad a range of uncertainty (figure #). This was also true when forecasting case counts at the one-week ahead horizon for both mean and median varieties of ensemble.

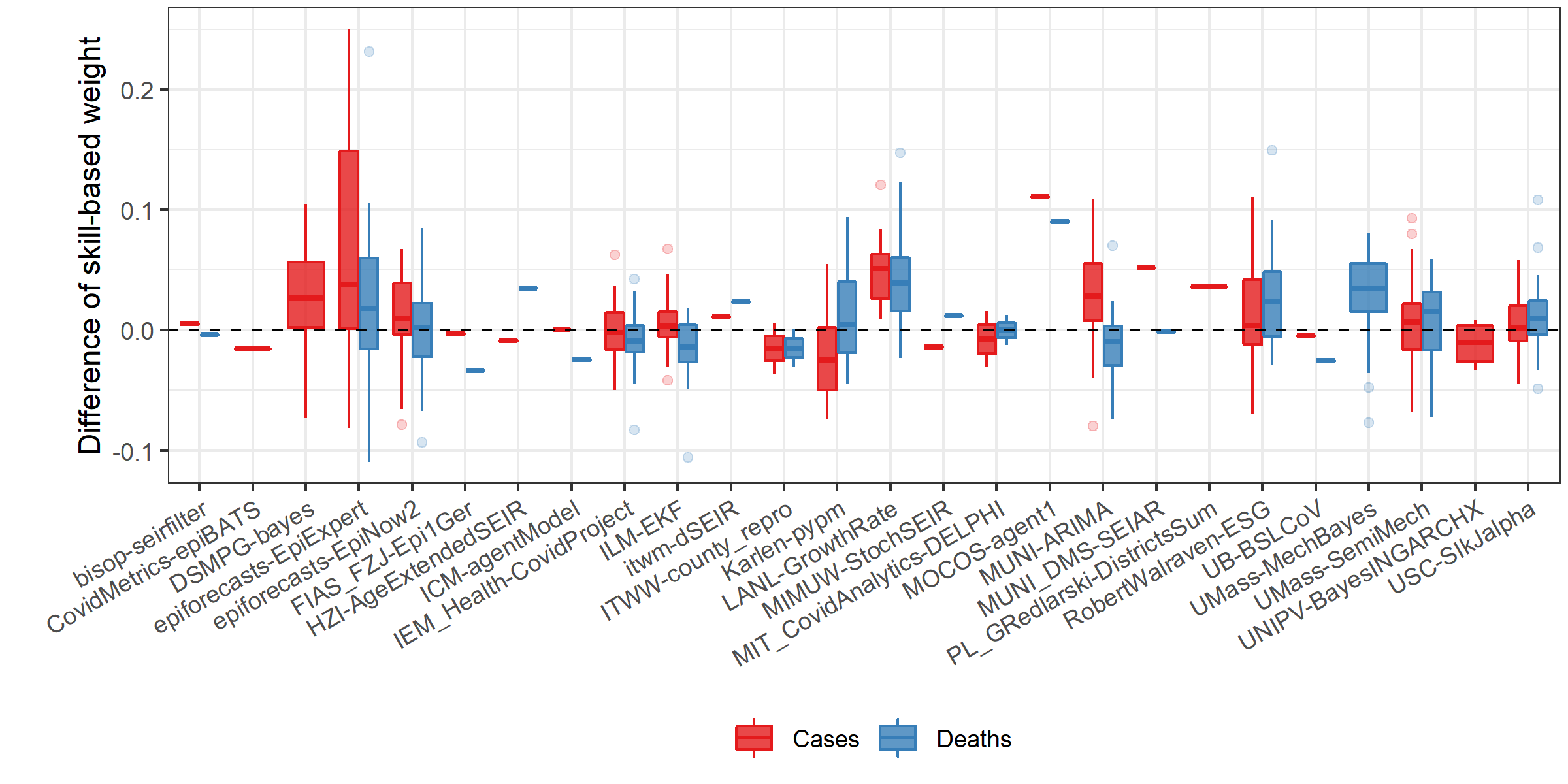
For both targets, the problem of underconfident prediction intervals reduced over longer horizons. This only improved forecasts of incident deaths, with the coverage of the 50% prediction interval the most accurate at four weeks. However case forecasts became overconfident in nearly all locations by the three and four week horizon.

*Averaging methods*

Ensemble forecasts that used any form of median outperformed the baseline model across all horizons and for each of 32 locations. The 766 ensembles using a median outperformed the baseline for 85% targets. Ensembles using the mean were less consistent in performance across countries, while still outperforming the baseline for 79% targets.

Forecasts for Iceland were a notable outlier, where mean forecasts failed to accurately forecast regardless of the type of mean ensemble used. This includes the worst of any ensemble performance, where the simple mean of forecasts for incident deaths in Iceland at three weeks ahead performed over seven times worse than the baseline forecast.

*Weighting methods*



*Figure #4: Change in model contribution of each model to ensemble forecasts: difference from weighting in a simple ensemble (dashed line, 0), to weighted by relative skill (averaged across all horizons), by target. Boxplot distribution represents average difference in model weight for each country, where data above the dashed line indicates that the model weight increased, and with anomalous differences shown as points.*

Weighting by skill created substantial differences in the contribution of individual models compared to weighting all models equally (figure #4), and these differences varied by target and over time (for more on relative performance of individual models, see forecast report). However, weighting by skill had relatively little impact on the performance of ensembles compared to simple equal weights. Of 256 targets, there was no greater than 2% difference in the number of ensemble forecasts that outperformed the baseline, between ensembles weighted by skill on average, by horizon, or a simple average for each of the mean and median approaches.