Proposed analysis plan

Rike Becker

January 5, 2024

1 Problem Statement

We want to consider how ensemble performance is related to ensemble size, i.e. the number of component models within the ensemble.
Main issue with 'straightforward' analysis: model availability fluctuates considerably throughout the time period under study:
(With 'straightforward' I mean: Choose entire time period or a large subset thereof, possibly filter out a small number of models, then generate all ensembles of size k by recombining from individual model set.)

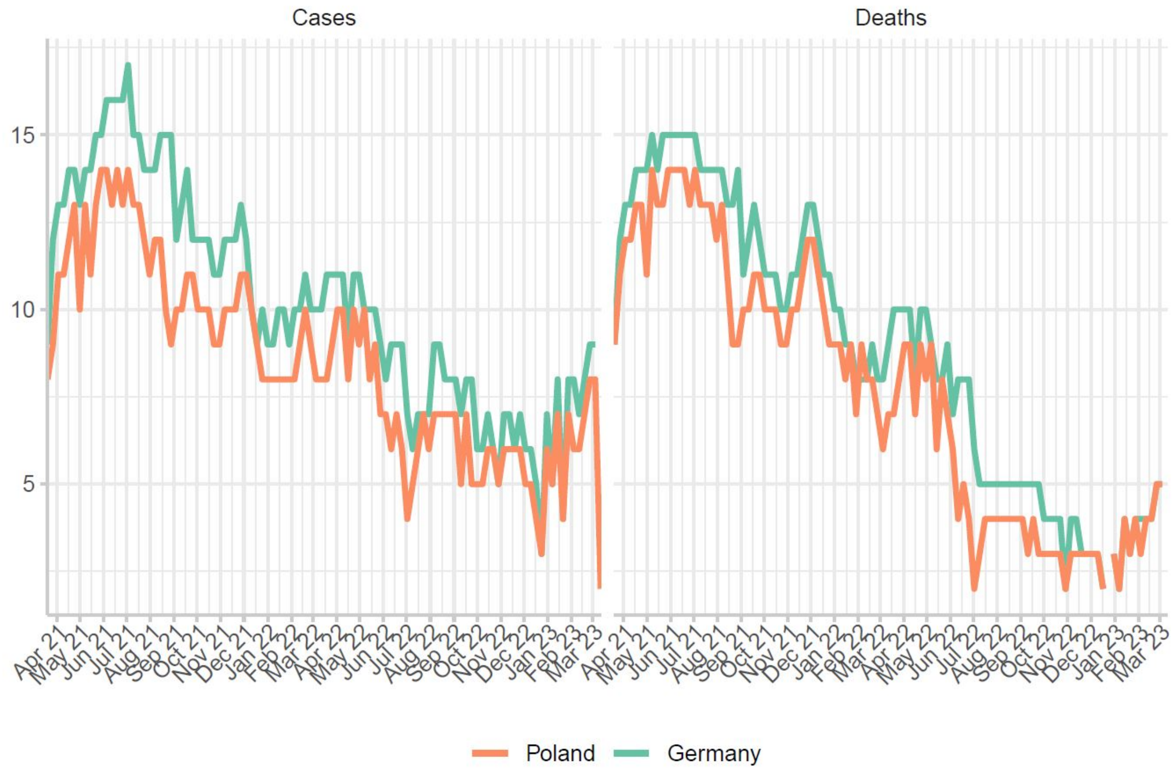

Figure 1: Caption

There doesn't seem to be a sufficiently large time period with somewhat stable participation that could be used for this. If we limit ourselves to only the models that are always available, we only have a small set available and can't vary k much.

2 Proposed Fix A: Split up time periods

The first proposed (and already tried) fix was to split up the entire study period into chunks, with the number of available component models then being somewhat constant within each time period.

The issue with splitting up time periods is that overall model or ensemble performance cannot be regarded as "identically distributed" (or something close to that) across time periods. For instance, during the summer months of 2021, where participation was quite high, performance was also overall better. This did not merely affect *absolute* performance, but also *relative* performance ("target was easy, so all models performed similarly"?).

Because of this and perhaps additional reasons, results (relative performance of size-k-ensemble) vary considerably and it is unclear if and how results could be standardized and thereby how to facilitate their communication. Some artefacts also suggest that periods may be too short for this analysis, i.e. variance might be too high.

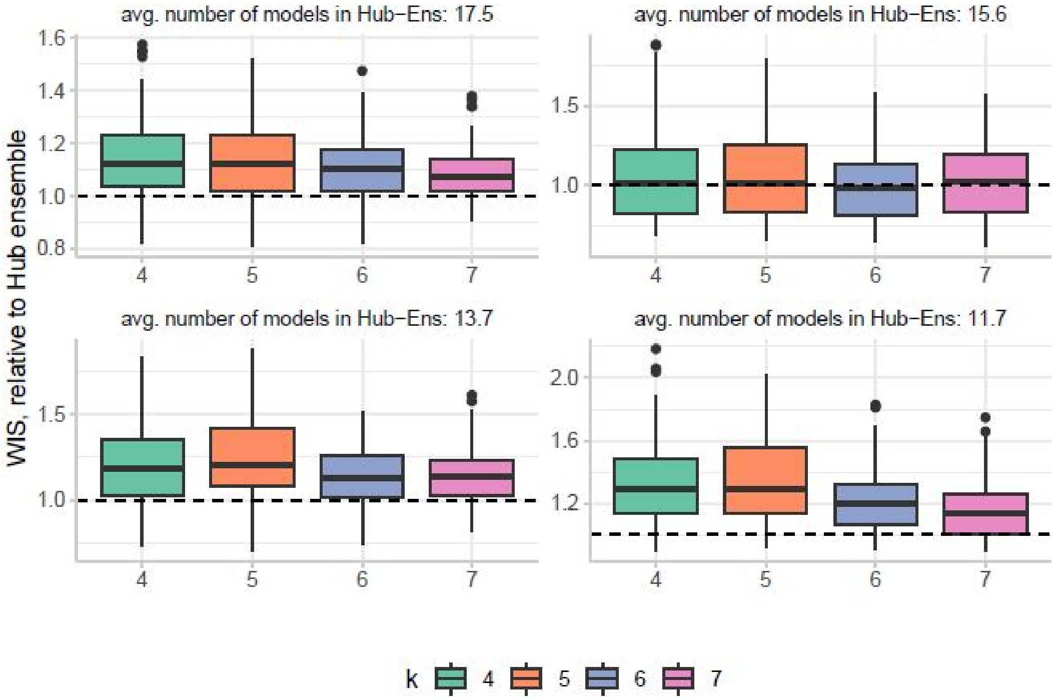

Figure 2: Caption

3 Proposed Fix B: Utilize pairwise comparisons

The new proposed analysis plan would look like the following:

Treat time period as a whole, that is, recombine from all available models across and build ensembles across the entire time, with some additional restriction parameters:

- q_m - individual model availability
  To be considered and included in the analysis, models must be fully[1] available for a proportion q_m of total forecast dates.
  Suggested value: $q_m = 0.5$

- q_e - availability of recombined ensemble by forecast date
  For an ensemble to be considered "available" at any given forecast date, a proportion of at least q_e of its component models must be available at that date. This means that we *do* allow for an ensemble of size k to have less than k models for some of the time. In theory, this seems to be a nice representation of a hub where members (to a certain degree) might drop in and out.
  Suggested value: $q_e = 0.75$ and possibly $q_e = 0.66$ for $k = 3$

- q_t - total availability of recombined ensemble
  Lastly, for an ensemble to be considered in the final scoring, it needs to be available (according to q_e-flag) for a proportion q_t of the time period.
  Suggested value: $q_t = 0.5$??? This really depends on pairwise comparisons

- Possibly, but not critically important: For each k, sample from ensembles if total number of ensembles is larger than for instance 350 or 500. This might however clash with reporting minimum/maximum performance. Upside: computation times :)

This means that we propose recombined ensembles from all models that "pass" q_m, then flag (0 or 1) each ensemble's availability by forecast date according to q_e and finally check if the resulting ensemble passes the total availability threshold q_t.

For scoring, to account for different availability of the recombined ensembles, we could use pairwise comparisons as in previous works of the US and European Covid-19 Forecast Hubs.

Similar to analysis of André Amaral (Nowcast Hub) and Spencer Fox (US CovidHub, ILI), one could finally highlight certain models/ensembles:

- baseline (constant across k)

- full hub ensemble (constant across k)

- "best model" ensemble (varies across k)

- "reliable" ensemble (ensemble of size $k = 5$ with perfect availability. constant across k). This could highlight what performance can be achieved

I think that this procedure of analysis would be more straightforward and communicable, and should account for data availability issues caused by the intermittent submission behavior.

[1] That is, all 23 quantiles across all 4 horizons