Abstract: We examine the relative performance of volunteer versus MTurker forecasters in the HFC forecasting project. A mid-trial-period decision to augment two of three initial experimental conditions with MTurk forecasters left the research design unable to estimate the effect of showing human forecasters data charts and/or machine forecasts on accuracy. We leverage the fact that only a portion of questions were captured by the data and machine forecast generation system to produce such estimates, and also examine how those estimates change over time in response to improvements in the quality of the data and machine forecasting system.

Funding acknowledgement: This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17071900005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Additional details:

This was the initial experimental design:

Condition	Treatment	Population
A	None; control	Volunteers and Turkers
B	Chart	Volunteeers
C	Chart and machine forecast	Volunteers and Turkers

Condition B has the lowest average Brier score; the other conditions are roughly on par, which superficially suggests that seeing a chart is beneficial, seeing a model forecast is not. The problem with this is that there are two treatments, whether a user sees a chart and whether a user sees the machine forecasts, as well as the confounder of whether a user was a volunteer or turker. With three groups and three variables there are not enough degrees of freedom to pin down the effect of any of these.

Condition	Has turkers?	Sees chart?	Sees models?
A	X		
B		X	
C	X	X	X

However, an additional factor is that we don't actually produce charts and machine forecasts for all IFPs. Only about one third of questions have chartable data and the possibility of a machine forecast. Thus there are actually really two sub-groups within each experimental condition: those looking a questions for which there is no data in the system, and those looking at a question for which there is a chart and machine forecast, although they may not see it.

Condition	Chartable data	Has turkers?	Sees chart?	Sees models?
A		X		
A	X	X		
B				
B	X		X	
C		X		
C	X	X	X	X

We can't just compare the two sub-groups in condition B that have/don't have a chart since the questions themselves might be more difficult for one type. But now there are 6 subgroups and 4 factors we would need to estimate: the chart and model effects, as well as the confounders turker and whether an IFP had chartable data.

In addition to estimating the general effects of seeing a chart and seeing a machine forecast on accuracy, when accounting for turker and IFP effects, we also want to take into consideration the fact that the early machine forecasts and to a lesser extent charts suffered from quality issues that improved over time. Thus we also want to estimate time-varying effects for those two factors. Preliminary results show small positive and negative relations for seeing charts and seeing bad machine forecasts on accuracy, but both outweighed by a much stronger negative turker effect.