Assessing Amazon Turker and automated machine forecasts in the Hybrid Forecasting Competition*

Andreas Beger[†] *and Michael D. Ward*[‡]

2018-12-12

Abstract

The Hybrid Forecasting Competition (HFC) is an ongoing project to develop hybrid geopolitical forecasting systems that combine human- and machine-generated forecasts. The first project trial period took place in 2018, during the course of which volunteer participants, Amazon Mechanical Turkers, and an automated time series forecasting module we developed provided forecasts for more than 150 questions covering a diverse set of topics. We assess several aspects of the accuracy of human and machine forecasts.

The Hybrid Forecasting Competition (HFC) is an IARPA program that seeks to develop methods for hybrid geopolitical forecasting system that combine human and machine forecasts to answer a broad range of questions about economic, political, health, and other events and trends. The first trial period, or RCT, took place in 2018, during the course of which hundreds of volunteer and Amazon Mechancial Turk forecasters, as well as automated machine models, answered more than 150 questions (IFPs) covering a broad range of issues.

We worked on one of the competition teams, and specifically by contributing a time series forecasting module. Among the questions we investigate in this paper are: (1) how does the accuracy of turker forecasters compare to volunteer forecasters, (2) what was the impact of time series charts and machine forecasts on human accuracy, (3) are there areas in which machine forecasts performed better than human forecasts, and vice versa, and (4) what explains the good performance of one particular group of volunteer forecasters who saw time series charts but were not exposed to model forecasts.

The Hybrid Forecasting Competition

The goal of the HFC is to find ways to optimally combine human and machine forecasts. For example, machine forecasts can be reliable and scalable, but are constrained by available data, idiosyncratic questions, and cold start problems when a corpus of historical data is not available. Human-generated forecasts on the other hand are more flexible, but also more costly to scale and subject to various cognitive biases (e.g. Kahneman and Egan 2011).

The competition is organized around Good Judgement-style forecasting tournaments (P. Tetlock 2005, P. E. Tetlock, Mellers, and Scoblic (2017)), where forecasters answer and are scored on a diverse pool of questions. Examples from the first trial period, RCT-A, include:

- What will be the long-term interest rate for South Africa (ZAF) in July 2018?
- How many deaths perpetrated by Boko Haram will the Council on Foreign Relations report for June 2018?

[*] **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.

Prepared for the 7th Annual Asian Political Methodology Conference, 5-6 January 2019, Kyoto, Japan

[†] Predictive Heuristics, adbeger@gmail.com

[‡] Predictive Heuristics, michael.don.ward@gmail.com

- What will be the daily closing spot price of Brent crude oil (USD per barrel) on 31 May 2018, according to the U.S. EIA?

Each question includes 2 to 5 answer options to which forecasters must assign weights summing to 1. Once a question has closed and a resolution is available, the forecasts are scored. The primary performance measures were various aggregations derived from multinomial Brier scores.

In addition to this basic "human forecaster" tournament, competitors were expected to implement features that would in some fashion or argument these human forecasts with machine-generated tools and forecasts. An explicitly requirement for the latter is that they are automated systems, e.g. an ad-hoc hand-tuned and expert-implemented machine model to forecast on a question would not be allowed, rather it would have to be a system that can generate such a model. On the other hand, if a forecaster has the skills and inclination to use data and model as a forecasting tool, that is perfectly allowed.

Automatic systems for acquiring data and producing time series charts and forecasts

One features of our team's approach was a system that could automatically associate some of the RCT-A questions with a clearly corresponding time series and display these to a forecaster. This is based on a data platform which automatically collects and updates data from a variety of sources, and which can associate questions, based on their title, to an appropriate transformation of a data set if it is in the platform and matches a known, pre-specified pattern or template. If data is found for a question, it can then be shown to users as a simple time series chart accompanying the relevant question.

Additionally, we could use the time series to generate a machine forecast based on a univariate ARIMA model, and which then would either be shown to a user, and/or submitted separately as a standalone forecast. The module generating these forecasts, "basil-ts", was based on the Auto ARIMA model in Hyndman and Khandakar (2008), which consists of a ARIMA-family model with an automated algorithm for determining a reasonable specific model structure, e.g. whether and how many differencing orders to apply, AR and MA orders, and some other parameters. This is wrapped in additional functionality needed to meet the automation requirements, e.g. recognizing how far a forecast needs to extend, data pre-processing, converting time series to answer option forecasts, updating forecasts with additional information if available, etc.

Research design for RCT-A

Since the scientific goal of HFC is to evaluate various hybrid forecasting techniques, the primary aspect of the research design for RCT-A was to assess what impact exposure to the time series charts and model forecasts would have on human forecaster accuracy. Incoming forecasters were assigned to one of three experimental conditions. The first group, A, served as control group and only had access to a basic version of the online platform showing the question information and tools to enter weights for the answer options. The second group, B, could also see the time series charts, and the third group, C, could see the chart and machine forecast. All forecasters, regardless of group, were forecasting on the same set of questions.

There were elements in the research design to assess other design choices but they are not relevant to the set of questions we seek to examine here, thus we will not discuss them.

Table 1: Summary of original research design.

Condition	Treatment
A	None; control
B	Chart
C	Chart and machine forecast

Comparison of groups A and B should have shown the effect of seeing a time series chart, and of groups B and C for the effect of seeing a model forecast. In practice, there were issues that complicated the effective design or forecaster groupings and treatments:

- **Amazon Mechanical Turk forecasters**. Due to lower than expected activity levels, turker forecasters started to be provided several weeks into the first trial period. Given activity levels at that time, a decision was made to assign turkers to either condition A or C, but not B. This meant that the group whose treatment was to only see charts (B) consisted only of volunteers, while the other two groups contained mixed populations of volunteers and turkers, thus adding a confounding factor for assessing the impact of charts and machine models.
- **Gaps in data and machine coverage**. Only about 1/3 of IFPs had time series data available, meaning that within each condition group, most questions did not have a chart nor model. There was also a small number of instances where data was available, but a machine forecast was not. One example were questions related to FluNet influenza case counts, where a change in the data source broke the ability to update chart data, which would have required models to forecast over excessive time horizons. Since the availability of data was related to the type of question, it is not possible to rule out the possibility that questions with data were systematically different in their difficulty from questions without data.
- **Changes in the data and machine forecasting platforms over time**. Both the data platform and the machine forecasting system experienced various bugs and related issues, especially during the earlier portions of RCT-A. These problems in some cases resulted in incorrectly aggregated or otherwise inaccurate data, insufficient updating which led to data in the platform falling behind source availability and thus requiring forecasts over longer time periods than necessary, and bugs in the machine forecaster that led to no or bad forecasts. As these were addressed over the course of time, the quality of the data and machine forecasts displayed to some users varied at any given point in time and IFP, in ways that are difficult to reconstruct in retrospective.

Table 2: Effective treatments by group after addition of turkers to conditions A and C.

Group	Condition	Forecaster	IFP_Group	Sees chart?	Sees model?
1	A: no chart	Turker	No TS data		
2	A: no chart	Turker	Chart only		
3	A: no chart	Turker	Chart and model		
4	A: no chart	Volunteer	No TS data		
5	A: no chart	Volunteer	Chart only		
6	A: no chart	Volunteer	Chart and model		
7	B: chart only	Volunteer	No TS data		
8	B: chart only	Volunteer	Chart only	X	
9	B: chart only	Volunteer	Chart and model	X	
10	C: chart and model	Turker	No TS data		
11	C: chart and model	Turker	Chart only	X	
12	C: chart and model	Turker	Chart and model	X	X
13	C: chart and model	Volunteer	No TS data		
14	C: chart and model	Volunteer	Chart only	X	
15	C: chart and model	Volunteer	Chart and model	X	X
16	Machine	Machine	Chart and model		

In respect to the first two problems, which we can quantify, the effective treatment groups were more complicated and are shown in Table 2. Within each of the original experimental conditions (A, B, C), there were potentially both volunteer and turker forecasters. They made forecasts on group of IFPs for which data was not available at all, on which data and a chart and a machine forecast were available, and also a small number of questions for which data and a chart, but not a machine forecast, were available. Since it is not possible to show a chart for an IFP for which data were not available, even in conditions B and C actual exposure to charts and model forecasts was also conditional on the IFP group in respect to data availability.

Table 3: Average Brier score by forecaster group

Forecaster	avg_Brier	n
Machine	0.39	1975
Turker	0.43	39140
Volunteer	0.32	7816

The last two columns thus mark the groups of forecasters who actually would have seen a chart or model forecast on a question. Finally, we also mark the subset of questions for which there was a machine forecast.

Data

The data that we will use for the empirical exploration consist of single forecasts that a user (volunteer, turker, machine) made at a particular time (date) for a question (IFP), i.e. a set of probabilities for the number of options that that question had.

RCT-A lasted from March 7th to September 7th 2018. However, turkers did not enter until May 2nd, and for an unrelated reason we also discard data after August 2nd. This leaves a total of 48,931 forecasts—7,816 from volunteer forecasters, 39,140 from turkers, and 1,975 from machine—for 156 IFPs, of which 101 did not have TS data, 6 had TS data but not a machine forecast, and 49 had both TS data and a machine forecast.

The basic measure of forecast quality are multinomial Brier scores that range from 0 to 2, with lower values indicating better performance. This is used for questions with only two or several unordered responses, and calculated as:

$$mBS = \sum_{i=1}^{R}(f_i - o_i)^2$$

In which R is the number of answer options, f is a vector of probabilities for each option, and o is a vector indicating the correct answer option. The score ranges from 0 to 2, lower values indicating better accuracy.

For questions with ordered answers, an ordinal variant of the multinomial Brier score is used so that "near-misses" are correctly penalized less than far misses, unlike in the multinomial Brier score. For a question with 4 answer options, it would be calculated with the following steps:

1. Split the ordinal categories (A-B-C-D) into cumulative binary pairs, aggregating the forecast probabilities for each grouping of categories (A-BCD; AB-CD; ABC-D).
2. Calculate the multinomial Brier score for each of the binary categories.
3. Average across the binary category scores to obtain the final Brier score.

The values of this score also range from 0 to 2.

The primary covariates or independent variables of interest correspond to the various factors defining the groups of forecasters in Table 2. We did not aim to incorporate other external factors, e.g. characteristics of the users beyond whether they were volunteers or turkers, for example.

Empirical analysis

Table 4 summarizes the average Brier scores for all 16 treatment/forecaster/IFP groups in Table 2. As it is quite unwieldy, we include it for reference and will discuss specific insights in more details below.

Table 4: Average Brier scores for all 15 distinct human IFP/treatment groups, as well as the machine forecasts

Group	Condition	Forecaster	IFP_Group	avg_Brier	sd_Brier	n
1	A: no chart	Turker	No TS data	0.45	0.50	6690
2	A: no chart	Turker	Chart only	0.55	0.56	304
3	A: no chart	Turker	Chart and model	0.39	0.40	2907
4	A: no chart	Volunteer	No TS data	0.31	0.54	510
5	A: no chart	Volunteer	Chart only	0.36	0.46	18
6	A: no chart	Volunteer	Chart and model	0.28	0.34	274
7	B: chart only	Volunteer	No TS data	0.26	0.48	1702
8	B: chart only	Volunteer	Chart only	0.42	0.42	57
9	B: chart only	Volunteer	Chart and model	0.23	0.31	973
10	C: chart and model	Turker	No TS data	0.45	0.50	19502
11	C: chart and model	Turker	Chart only	0.56	0.52	990
12	C: chart and model	Turker	Chart and model	0.38	0.40	8747
13	C: chart and model	Volunteer	No TS data	0.38	0.60	3090
14	C: chart and model	Volunteer	Chart only	0.59	0.57	73
15	C: chart and model	Volunteer	Chart and model	0.32	0.40	1119
16	Machine	Machine	Chart and model	0.39	0.36	1975

Table 5: Average Brier score by forecaster group

Forecaster	A: no chart	B: chart only	C: chart and model	Machine
Machine				0.39
Turker	0.44		0.43	
Volunteer	0.30	0.25	0.37	

Aggregate results

Table 5 shows a high level summary of average Brier scores by forecaster type and condition. Overall, volunteer forecasters in condition B had the best accuracy, turkers overall appear to have had poorer accuracy, and the machine forecasts were somewhat middling and generally lower than expected, but without a doubt not competitive with volunteer forecasts, even if they were forecasting without the benefit of a time series chart or model forecast.

These aggregate patterns don't reflect confounding factors, e.g. the possibility that turker accuracy is somehow influenced by the fact that they were only present in conditions A and C, etc. The next sections will examine in more depths these areas:

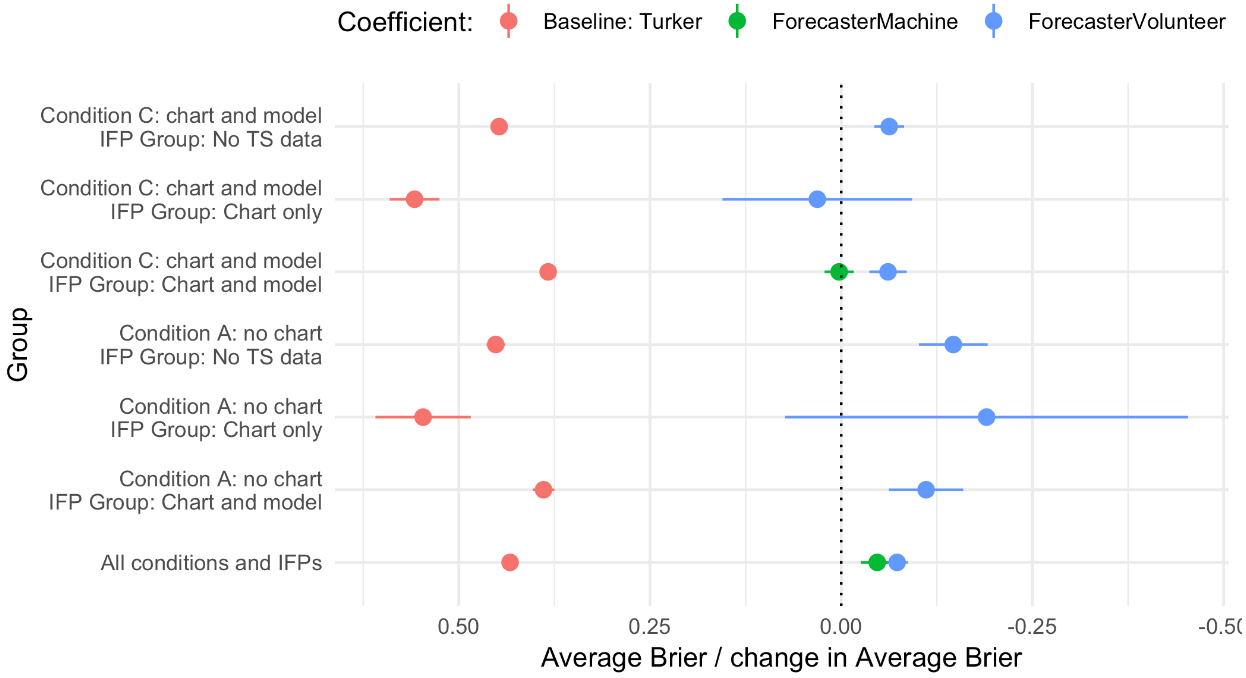
- Turker performance relative to volunteer forecasters
- What were the direct chart and model effects on the quality of a human forecasts?
- Are there specific areas in which machine forecasts seemed to outperform human forecasters?
- Why did condition B volunteers do so well compared to other forecaster-condition groups?

Turkers generally did worse

Turkers overall had worse forecast accuracy than either the machine models or volunteers. To rule a spurious relationship, we can compare the accuracy in directly comparable pairs, e.g. rows 1 and 4 in Table 4 consists of the average Brier scores for questions without TS data and in condition A treatment group for turkers and volunteers respectively, and show a difference of 0.14.

Figure shows the results of pairwise comparisons of the average Brier scores for all groups relative to Turker

Figure 1: Pairwise comparisons of turker versus machine and volunteer forecast accuracy. Each row shows estimates from a linear regression model with turkers as reference group, for different groupings of the experimental conditions and IFP data/model availability. The turker estimates (red dots) correspond to the average Brier for turker forecasters, while the other estimates correspond to the relative difference from the turker average for other forecasters. Points further to the right correspond to lower Brier scores, i.e. better performance. Volunteers on average performed better than turker forecasters in all conditions and question groups, except on the small number of charts-only IFPs, where estimates are uncertain.



forecasters. Each row in the plot corresponds to a linear model for forecasters from the corresponding condition and IFP group, and the pointranges show coefficient estimates and 95% confidence intervals for Turkers, Machine, and Volunteer forecasters. The Turker values correspond to the average Brier for them in this grouping, while the Machine and Volunteer estimates indicate the *relative* performance. I.e. for them, values below zero indicate an improvement. Note that the x-scale is reversed, so that better is further to the right.

In most groups volunteers have an advantage sufficient to generate p-values below 0.05, although the effects are minute in comparison to general between-forecaster differences: all models have very small adjusted R^2 values, below 0.01. Thus, while turker forecasters on average did not perform as well as volunteer forecasters, the difference in forecast quality for specific users and questions differs far more than the means between the two groups do.

Chart and model effects

Teasing out the chart and model effects is a bit harder through pairwise comparison since a forecaster seeing a machine forecast would have also always seen a chart. Instead, we directly coded whether a forecaster saw a chart or model forecast. Forecasters in conditions B and C and for questions that had a chart saw a chart, and forecasters in condition C looking at a question which had a machine forecast will have seen a model forecast. We also leave out the machine forecasts from the data for this portion, leaving almost 47,000 forecasts, of which about 25% were made with a chart available, and 21% also had a model forecast available.

Table 6: Model summary statistics

Model	Statistic	Value
Model 1: Linear model	N	46956
	R^2	0.01
	Adj. R^2	0.01
Model 2: Linear model with IFP random intercepts	N	46956
	Marginal R^2	0.01
	Conditional R^2	0.28

We estimated two linear models to predict Brier scores for a forecast. The first has the specification:

$$\begin{aligned}\text{Brier} = \; & \alpha + \beta_1 \text{SeesModel} + \beta_2 \text{SeesChart} + \beta_3 \text{IFPGroupChartOnly} + \beta_4 \text{IFPGroupChartAndModel} \\ & + \beta_5 \text{ForecasterVolunteer}\end{aligned}$$

All variables are dummy terms. The intercept α corresponds to the average Brier score for a Turker forecasting on an IFP that did not have chartable data, and who neither saw a chart nor machine forecast.

The second model includes random intercepts for IFP to account for the possibility that seeing charts or models may make users differentially willing to forecast on harder IFPs, i.e. those with higher average Brier scores. Otherwise the specification of covariates is the same.

$$\begin{aligned}\text{Brier} = \; & \alpha_{\text{Global}} + \alpha_{\text{IFP}} + \beta_1 \text{SeesModel} + \beta_2 \text{SeesChart} + \beta_3 \text{IFPGroupChartOnly} \\ & + \beta_4 \text{IFPGroupChartAndModel} + \beta_5 \text{ForecasterVolunteer}\end{aligned}$$

Figure 2 plots the coefficient estimates from both models. The results are consistent. Seeing a chart slightly helps, seeing a model slightly hurts forecast performance. However, the effects are small and not as important as whether a forecaster was a volunteer or turker. In addition, the models themselves capture only a small portion of the variation in Brier scores overall. Table 6 shows R^2 statistics for both models. The covariates account for less than 0.01 of the variation in Brier scores; the inclusion of random IFP intercepts in model 2 increases this to 0.28 (conditional R^2), but the fraction explained by the covariates together is still 0.01 (marginal R^2). Most of the variation remains either between forecasters or due to other random factors.

Thus, while we find that exposure to charts and model forecasts has a small impact on forecast quality, the effect sizes are small in comparison to other systematic factors in the model, and small relative to unsystematic and random factors not in the model. There is far more variation between individual forecasts (users, questions) than between the forecasts of users who saw a chart and/or model.

Models generally were disappointing, but did they do well in certain areas?

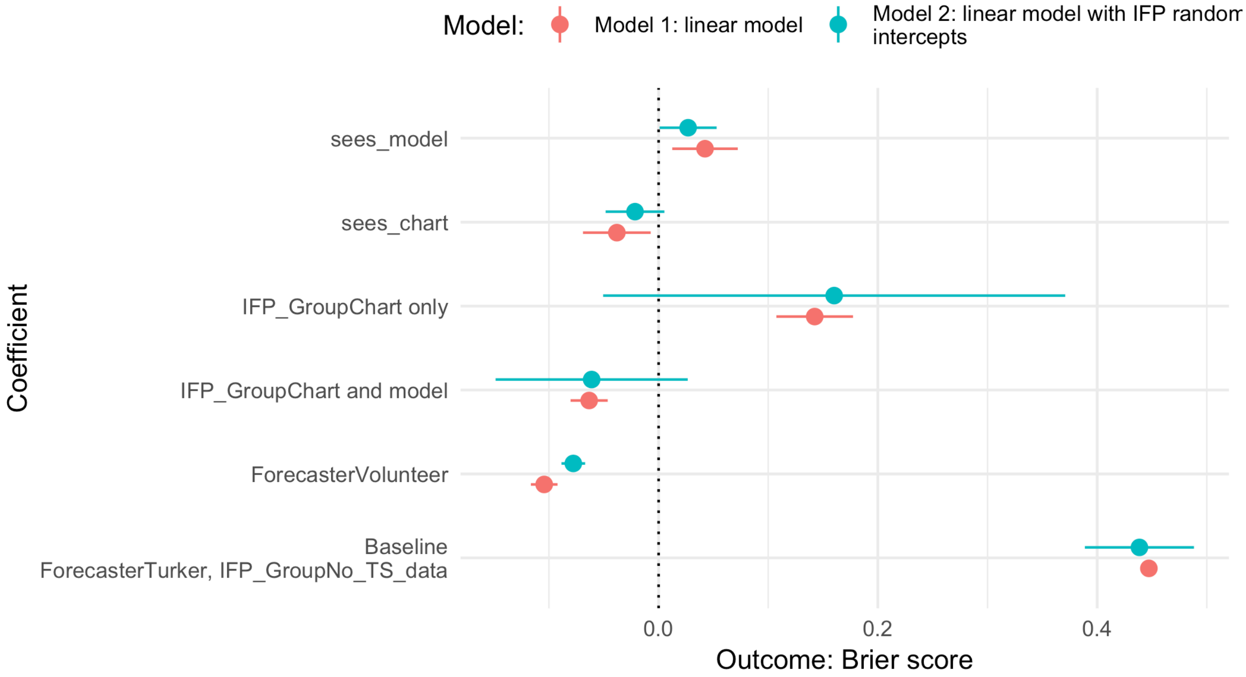
Although the machine models overall underperformed, it is possible that they performed well relative to human forecasters in certain areas. To examine this we looked at all forecasts for questions that had a machine forecast. This leaves us with 15,995 forecasts, approximately 2,000 of which were machine and volunteer forecasts, respectively, and the remaining 12,000 from turkers.

To estimate the relative performance of different forecasters, we again estimated a simple linear model, similar to those used for the turker pairwise comparisons. The model has the general form:

$$\text{Brier} = \; \beta_i \text{DataSource} + \beta'_{ij} (\text{DataSource} \times \text{Forecaster})$$

Data source is a discrete variable with $i = 14$ levels for the different data sources reflected in the set of IFPs. Based on previous exploratory analysis, we separated one data source, ACLED questions, by whether

Figure 2: Model estimates for the direct effects of seeing a chart and machine forecast. Model 1 is a linear model with the charted terms, model 2 in addition has random intercepts for each question (IFP) to account for question difficulty.



a question was binary with two answer options ("yes"/"no"), or not (3 or 5 options). Forecaster is also a discrete variable with 3 levels, but we leave out "Machine" as reference category, giving us $j = 2$ different levels. Thus, each of the β coefficients corresponds to the average Brier score for machine forecasts for that group of IFPs, and each β_i' coefficient is the relative difference of the average volunteer and turker Brier score compared to the corresponding baseline in β_i.

Figure 3 shows the coefficient estimates. The left panel shows the baseline machine forecast performance. The right panel shows the relative difference in average Brier scores for turker and volunteer forecasts; points to the left of the dotted 0 line indicate that volunteer or turker forecasters on average outperformed the machine forecasts.

The machine forecasts did better than human forecasts on ACLED 5-option, Boko Haram killings ("nigeria-security"), and hacking ("privacyrights") questions. All three data sources consist of count data derived from original event datasets, i.e. they require data aggregation before one can see a time series relevant to the question at hand, which might be a factor in the relative performance on these questions. On the other hand, human forecasters consistently did better on ACLED 2-option, food price index ("FAO-FPI"), financial ("fred"), and oil production ("opec") questions. The results for ACLED 2-option questions might be influenced by the particular mechanism by which time series forecasts were converted to answer option probabilities, which may explain the inconsistency with ACLED 5-option questions.

There are some hard to quantify factors that may explain some of the variation in human to machine performance:

1. As mentioned, some data sources (ACLED, privacyrights, nigeria-security) required relatively complicated data aggregation to derive usable count series, and which therefore may not have been accessible to all human forecasters.
2. Not all data sources are equally easy to automatically download and update, and at least in some of the data sources it was the case that models had to forecast over much longer time horizons due to

Figure 3: Linear model estimates for the relative performance of machine to volunteer and turker forecasts by data source. The left panel shows the baseline machine performance for each data source, the right panel shows the relative difference of the volunteer and turker forecasts to the corresponding baseline. The outcome variable is a forecasts's Brier score.

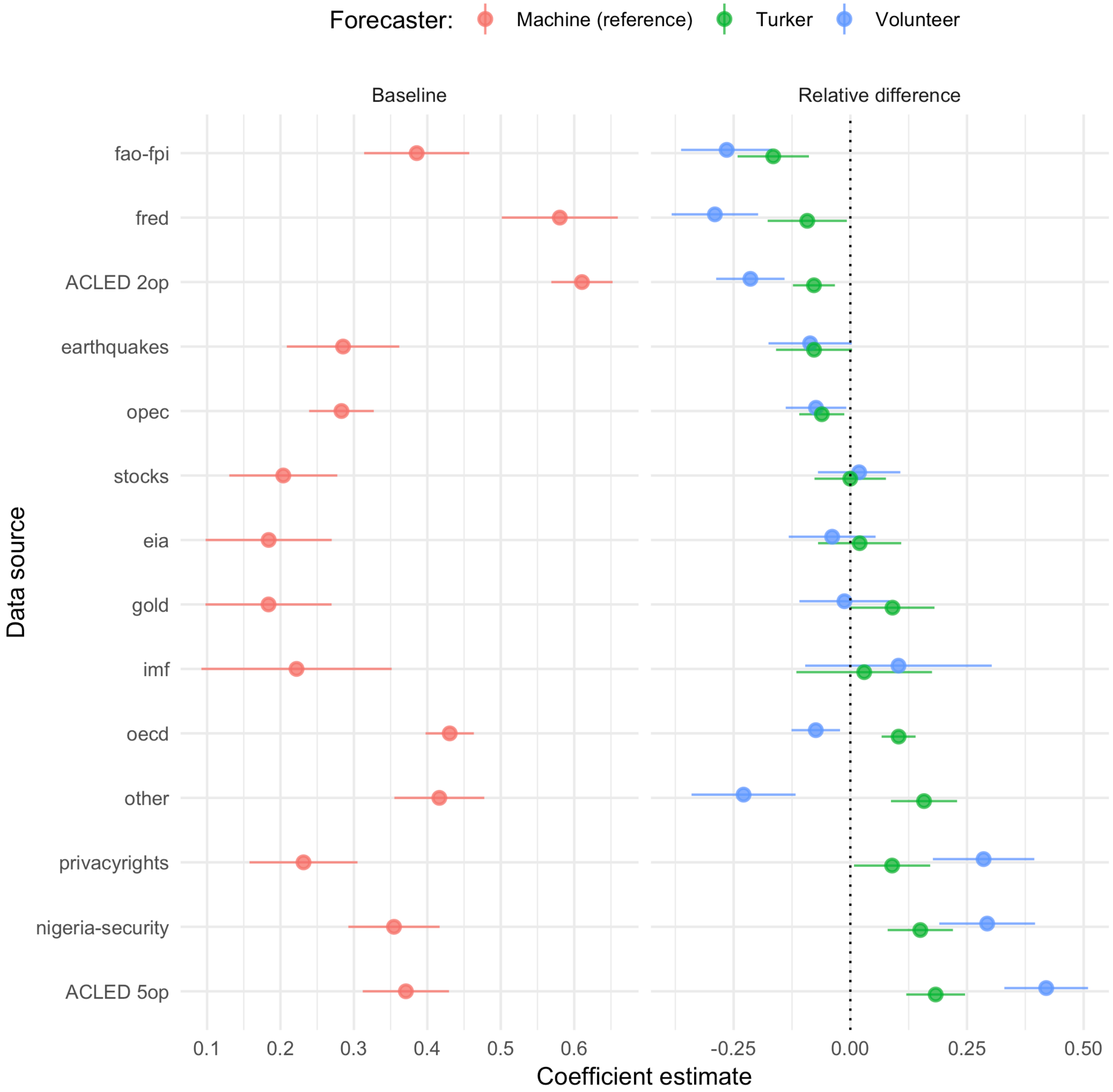


Table 7: Condition B advantage

Group	Condition	IFP_Group	avg_Brier	sd_Brier	n
7	B: chart only	No TS data	0.26	0.48	1702
4	A: no chart	No TS data	0.31	0.54	510
13	C: chart and model	No TS data	0.38	0.60	3090
8	B: chart only	Chart only	0.42	0.42	57
5	A: no chart	Chart only	0.36	0.46	18
14	C: chart and model	Chart only	0.59	0.57	73
9	B: chart only	Chart and model	0.23	0.31	973
6	A: no chart	Chart and model	0.28	0.34	274
15	C: chart and model	Chart and model	0.32	0.40	1119

outdated data, while a human forecaster would have been able to access more recent data.

3. Due to obstacles in data acquisition, in at least one case–OPEC oil production rates–a secondary and less accurate source was substituted. OPEC production values are delivered in monthly PDF reports that are difficult to automatically extract data from, and the alternative data source in some instances significantly deviated from the OPEC values. Thus models for these questions would have been forecasting with incorrect input data.
4. For questions in which the data source is at a monthly (or higher) resolution, humans have an advantage over univariate time series models in that they can incorporate sub-monthly information into their forecast. For many of these questions, if they were open for a period of one or two months, there would in fact have only been one or two distinct time series forecasts (barring practical live system issues and bugs).

A caveat to these trends is that the number of IFPs and forecasts for each data source are small and the results may thus be due to random variation. Results from future RCTs should test this.

Why did condition B volunteers do so well?

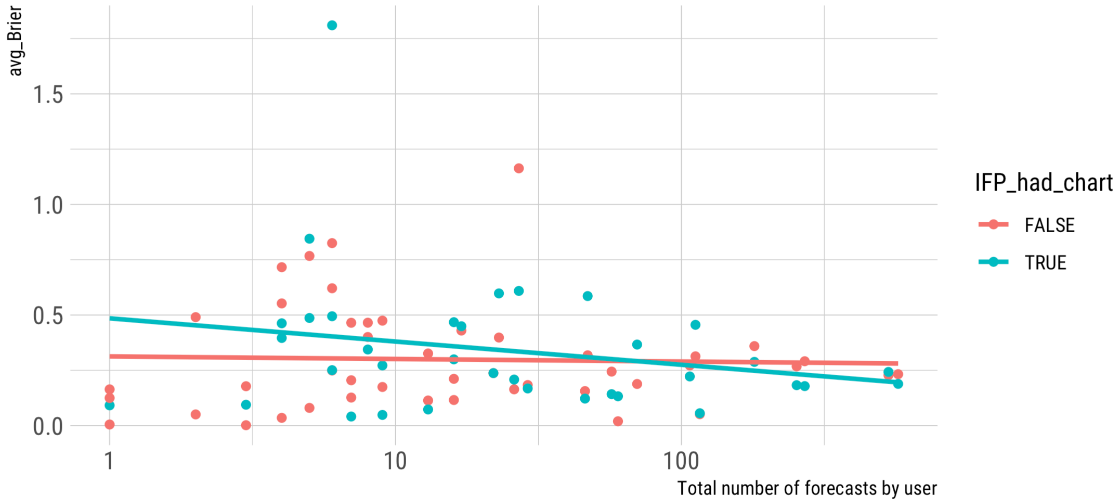
Condition B volunteers had the best overall performance out of the 16 groups shown in Table 4. Considering that the machine forecasts were of middling quality, one possible explanation is that the charts overall had a positive impact on human forecast accuracy, while seeing machine forecasts had a negative impact.

However, there are several additional patterns in the forecasts that are either inconsistent with this simple explanation or at least suggest the need for a more complex alternative. The first is that condition B forecasters did not only do better than other groups on questions that had a chart, but also on questions that did not have a chart. Table 7 shows the relevant rows from the average Brier scores for all forecaster groups in Table 4. With the partial exception of the few chart-only questions, condition B forecasters' advantage on questions without a chart is as pronounced as their advantage in questions that did have a chart. Considering that there were two to three times as many forecasts on questions without chartable data, most of the condition B volunteer advantage is driven by doing better on questions that did not have a chart.

Maybe there is, in addition to the direct beneficial effect of seeing a chart for a question, an indirect spillover effect whereby exposure to charts also improves a volunteer forecaster's ability to forecast on other questions without charts. For example, it could be that seeing charts is an effective reminder of taking relevant baselines into account when forecasting. There are still some curiosities:

There appears to be no chart spillover effect for condition C human and volunteer forecasters. Tabe 4 shows that condition C volunteers did worse on questions without time-series data than the condition A control group. Condition C turkers also did not do better than their condition A control group counterparts. Unlike condition B forecasters, they were exposed to machine forecasts as well, so maybe there is an equal

Figure 4: Average Brier for condition B volunteers by user and by whether a IFP had a chart. Intensive users are slightly more accurate than novice users on questions with a chart, but no more accurate on questions without a chart.



but opposite negative machine spillover effect.

Condition B volunteers who forecast a lot do no better on non-chart questions than novices. Figure 4 show average Brier scores as a function of the total number of forecasts a user has. The data consist of only forecasts from condition B volunteers. Each point is the average Brier score for a user's forecasts on the set of questions that did (blue) or did not have a chart (red). Novice users tend to be as accurate as intensive users on non-chart questions. This implies that the condition B volunteer advantage on non-chart questions is not related to how active a user was. If there is a spillover effect, even very inactive users seem to somehow benefit from it.

Condition C volunteers do well on questions where we know that the data displayed was wrong. There were several instances where we know that the data displayed in a chart were inaccurate. Figure 5 shows the first example, oil production figures for Iraq. The main chart shows several time series depicting oil production in Iraq, according to different sources. The canonical data are contained in monthly OPEC oil market reports, each of which contains production figures for the past 3 months. Those are shown in the sequence of short time series on the mid-right part of the main chart–the figures are somewhat inconsistent between reports, but this is another peculiarity. Because this data is difficult to obtain programmatically, an alternative data source is used instead. This is what would have been shown in the charts, and what informed the machine forecast. It is incorrect data. The dotted lines show the values separating the answer options for this question. The chart placed the production values in an entirely incorrect answer category.

Curiously, the condition B volunteers performed better on this question than any other forecaster, with an average Brier score of 0.07, compared to 0.11 and 0.45 for volunteer forecasts in conditions A and C (Table 8). The number of forecasts here are low however, and only the last three groups (machine, turkers in condition A, volunteers in condition C) have pairwise statistically significant differences in average Brier score from the condition B volunteers.

Figure 5: Iraq oil production by data source. Official figures are to be drawn from monthly OPEC reports, each of which includes values for the past 3 months. Due to difficulty in automatically processing the PDF reports, an alternative data source was used in the chart and to create the machine forecast.

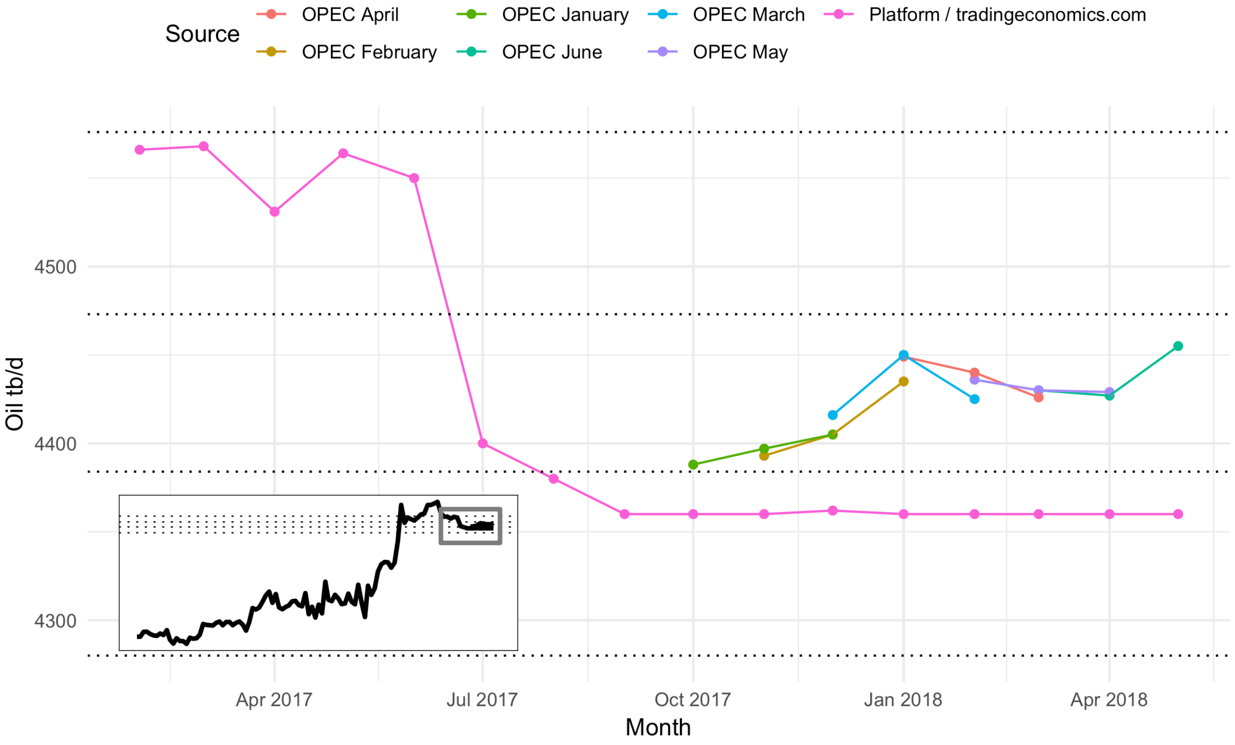


Table 8: Forecaster performance for “How much crude oil will Iraq produce in May 2018?”

Forecaster	Condition	avg_Brier	N
Volunteer	b	0.07	5
Volunteer	a	0.11	5
Turker	c	0.18	87
Machine	n/a	0.25	14
Turker	a	0.35	23
Volunteer	c	0.45	21

Another example with incorrect data is an early ICEWS question that fell outside the common period and thus did not have forecasts from turkers. Performance values for this question are shown in Table 9. Condition B volunteers also outperformed all other groups on this question, although the pairwise differences with the other groups are not statistically significant.

Table 9: Forecaster performance for “How many material conflict events involving Occupied Palestinian Territory will ICEWS record in March 2018?”

Forecaster	condition	avg_Brier	n
Volunteer	b	0.92	14
Machine	n/a	1.00	18
Volunteer	a	1.11	7
Volunteer	c	1.11	25

Even when the machine predictions were good, condition B volunteers did better than their C counterparts. If bad machine predictions influence human forecasts and are the reason why the beneficial chart effect is only apparent in condition B forecasters, and not condition C forecasters, then it should also be the case that condition C forecasters do well on those questions where the machine forecasts were good. This is no the case. Table 10 summarizes performance for forecasts on the 10 questions where the average machine forecast performance was better than the average performance for all other groups of forecasts, and where all groups of forecasters were represented, i.e. had forecasted. Since forecasters forecasted a differential amount of times in a way that is related to question difficulty, we cannot summarize the performance through a simple average. The relative performance measure shown in the table is instead calculated by first averaging Brier scores for each group on each question, calculating each group’s relative performance as percentage deviation from the overall average of group averages for that question, and then finally averaging these per-question relative performances over questions for a final measure for each forecaster group.

Table 10: Performance for questions where the machine forecasts had the lowest average Brier score accross groups. The 'avg_rel_Brier' column shows the expected deviation of a group's forecasts from the overall question average between groups, in %. Negative values indicate that a group's forecasts tended to do better than the overall average performance on a question. Had condition C forecasters sufficiently adjusted their forecasts in the direction of the machine forecasts, they should have outperformed other comparable forecaster groups.

Forecaster	Condition	avg_rel_Brier (%)	N_users	N_fcasts
Machine	Machine	-0.55	1	423
Turker	A: no chart	0.23	152	610
Turker	C: chart and model	0.25	469	1890
Volunteer	A: no chart	-0.13	8	50
Volunteer	B: chart only	-0.04	20	216
Volunteer	C: chart and model	0.24	40	224

The forecasts summarized in the table are for the small set of questions where the machine forecasts did very well, with Brier scores that were 55% less than the between group average for a question. If human forecasters who see machine forecasts align their forecasts towards the machine forecasts, they should have done really well here. But they don't. The condition B volunteers still outperform their "i saw a really good machine forecast" condition C counterparts, and to a lesser extend the same trend is apparent with turkers who saw neither chart nor model and those that did.

Each single example of an apparent consistency itself is based on small findings in noisy data. All of the effects—for turkers versus other forecasts, for the disappointing performance of the machine forecasts, for conditions or different groupings of IFPs—pale in comparison to differences in accuracy between individual forecasters and between forecasts for different IFPs. But on the other hand there are several inconsistencies so it also is a stretch to write all of them off to random coincidence. Condition B volunteers clearly did better than all other forecasters/condition groups, but it is not clear why.

Conclusion

We reported an exploratory analysis that examined several aspects of the performance of volunteer, Amazon Mechanical Turker, and machine-generated forecasts for a portion of the first trial period in the Hybrid Forecasting Competition (HFC) project. These initial findings suggest that:

- Amazon Mechanical Turker forecasters underperform compared to volunteer forecasters. It may still be that there is a subset of turker forecasters that manage to consistently outperform the average volunteer forecaster, a question we did not examine but which is plausible due to the overall small difference in volunteer and turker average performance.
- Volunteer forecasters who were able to see charts for the subset of questions with clearly associated time series overall had the best accuracy, even on questions that did not have chartable data. However, there are several potential empirical inconsistencies with the implied causal mechanism (charts have a direct beneficial effect on the question at hand, and some form of indirect beneficial effect on other, non-chart, question as well).
- Machine forecasts on average did not perform as well as volunteer forecasts, but slightly better than turker forecasts. They appear to have had an edge on questions with count series derived by aggregating event data on protests or violence (ACLED; Boko Haram violence in Nigeria) as well as hacking (Privacyrights hacking incidents). One commonality in these sources is that they require data aggregation in a way that is not easily replicable by a user without some moderate level of technical skill.

Additional information

Table 11: Average Brier scores for machine and other forecasts by question data source.

Data_source	N_IFPs	Machine	Turker	Volunteer
ACLED 2op	7	0.611	0.533	0.396
ACLED 5op	4	0.371	0.554	0.790
earthquakes	2	0.285	0.207	0.199
eia	5	0.184	0.204	0.145
fao-fpi	3	0.386	0.220	0.121
fred	2	0.580	0.488	0.290
gold	3	0.184	0.274	0.171
imf	1	0.222	0.252	0.325
nigeria-security	2	0.355	0.504	0.648
oecd	7	0.430	0.534	0.356
opec	5	0.283	0.222	0.209
other	2	0.416	0.574	0.188
privacyrights	2	0.231	0.321	0.517
stocks	4	0.204	0.204	0.223

References

Hyndman, Rob J., and Yeasmin Khandakar. 2008. "Automatic Time Series for Forecasting: The Forecast Package for R." *Journal of Statistical Software* 27 (3). Monash University, Department of Econometrics; Business Statistics.

Kahneman, Daniel, and Patrick Egan. 2011. *Thinking, Fast and Slow.* Vol. 1. Farrar, Straus; Giroux New York.

Tetlock, Philip. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

Tetlock, Philip E, Barbara A Mellers, and J Peter Scoblic. 2017. "Bringing Probability Judgments into Policy Debates via Forecasting Tournaments." *Science* 355 (6324). American Association for the Advancement of Science: 481–83.