Assessing forecast accuracy in the face of varying forecast difficulty*

James R. Bland†

February 22, 2026

Abstract

I propose a model for assessing forecasters' accuracy when raw forecast errors may be confounded by forecast difficulty varying over time.

1 Introduction

How can we assess the accuracy of forecasters when the difficulty of forecasting a variable is always changing? In some time periods, it may be relatively easy to forecast a variable, but in others it may not. Furthermore, if forecasters do not make forecasts in every time period, we may simply be ranking forecasters based on which ones selected into and out of the sample at convenient times. This paper proposes a Bayesian hierarchical approach to the problem, with forecast difficulty modeled as a latent process.

Using the Survey of Professional Forecasters (SPF), I compare one-quarter-ahead forecasts of the US unemployment rate to estimates published in the Federal Reserve Economic Data. One modeling hurdle to overcome is that forecasters mostly report their forecast in increments of tenths of a percentage point. While this amount of coarsening may not be substantial when comparing forecast *levels*, when comparing forecast *errors* one tenth of a percentage point can be large. As such, I account for this censoring by treating the rounded forecasts as interval-valued variables.

*This project benefitted greatly from comments by Rachel Leah Childers.

†The University of Toledo, 2801 Bancroft St, Toledo, OH 43606, USA, james.bland@utoledo.edu, orcid.org/0000-0002-7117-9998

The hierarchical approach assumes that each forecaster has access to a noisy signal of the forecast variable, and reports a censoring of this signal. The signal technology varies between forecasters and is estimated in the model, so the model can rank forecasters based on the estimated bias and precision of their forecasts.

While a ranking of forecasters does emerge from the model, I find that there is much posterior uncertainty in this ranking. This could be due to forecasters being similarly noisy in their forecasts. Importantly, the model's ranking is *not* just a replication of the ranking of (say) raw forecast errors. In fact, there is no noticable relationship between the model's ranking and a ranking based on these raw errors. If one then believes the model's correction to varying forecast difficulty, this suggests that raw forecast errors are poor measures of forecasters' accuracy.

One product of the model that may be of interest is its estimate of forecast difficulty, which is modeled as an autoregressive process. Unsurprisingly, forecast difficulty was estimated to be greatest during the COVID19 pandemic, and was generally greater during recessions than not during recessions.

The remainder of this paper is organized as follows. Section 2 introduces the two data sources used and some of their perculiarities. Section 3 discusses the model and estimation strategy. Section 4 presents the results. Finally, Section 5 concludes.

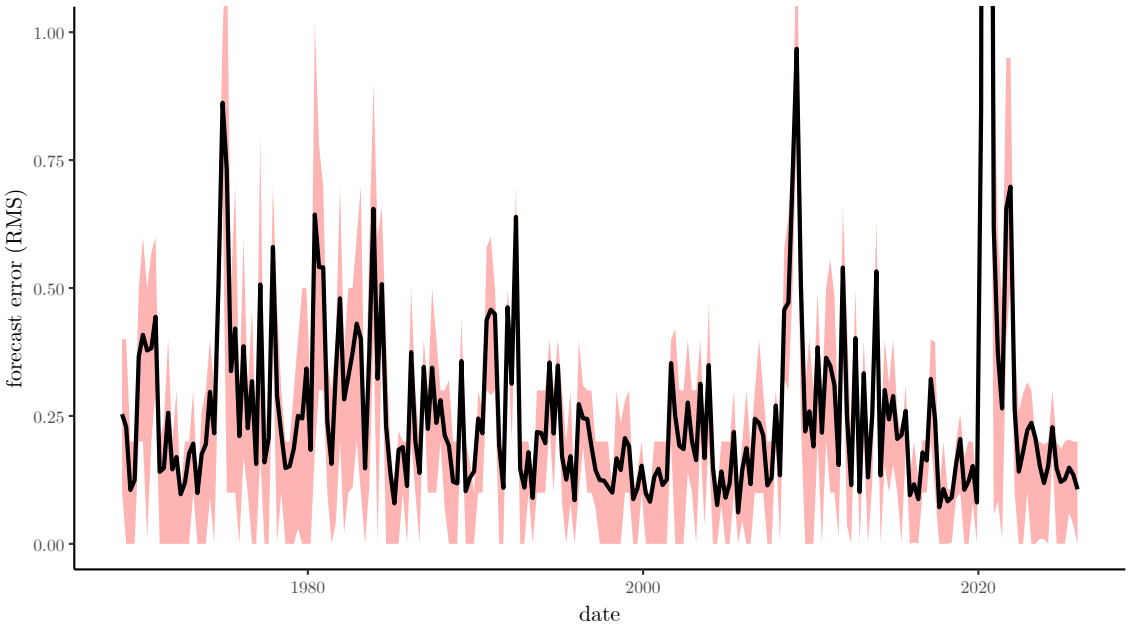
2 Data

I use data from two publicly available sources. First, I use the Individual Forecasts for the Survey of Professional Forecasters (Federal Reserve Bank of Philadelphia, 2026). I extract the "UNEMP2" variable, which is a one-quarter-ahead point forecast of the US unemployment rate, measured as a percentage. At the time of downloading (2026-02-21, last updated 2025-11-17), this dataset contained 9,178 forecasts from 462 forecasters, going back to 1968Q4. I drop all forecasters who made fewer than twenty forecasts. This means that the dataset used for estimation contains 7,258 forecasts and 151 forecasters.

I compare these forecasts to the estimated unemployment rate published by the St. Louis Federal Reserve (Bureau of Labor Statistics, 2026). Specifically, I use the series "UNRATE". Figure 1 plots the root-mean-squared forecast error for the sample. Here we can see that average forecast errors in magnitude of about 0.1-0.3 were common, but they rearely became larger than 0.5. The main exception to this was during the 2020 COVID19 pandemic, where the forecast error is (literally) off the chart.

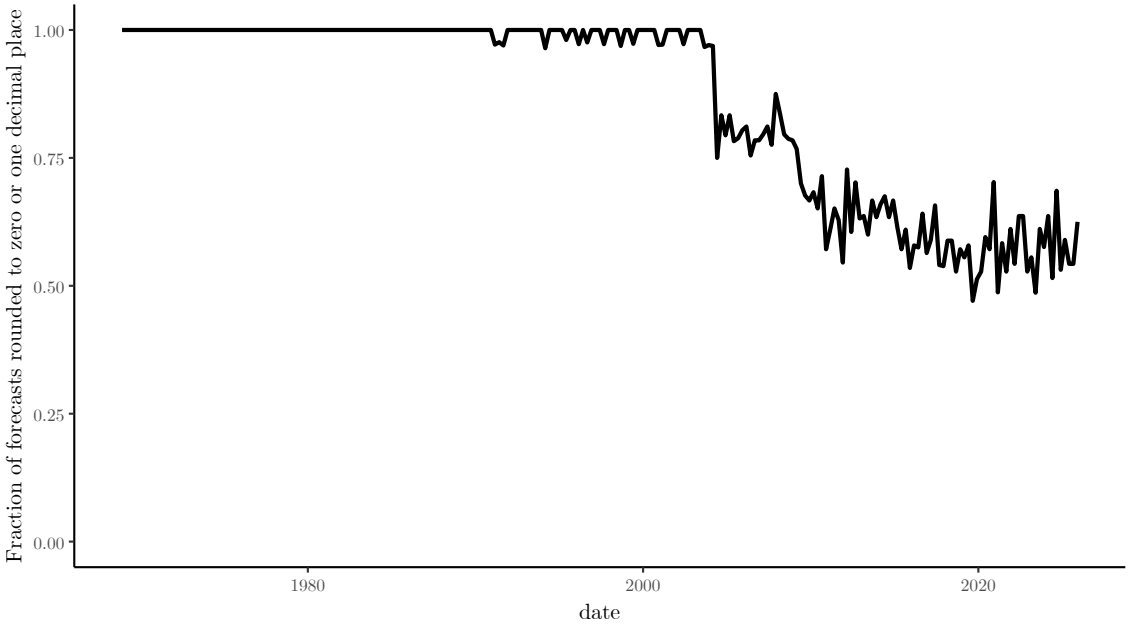
Figure 2 shows an important quirk of the forecasts. In particular, it shows the fraction of forecasts that are reported as either an integer percentage or a percentage

Figure 1: Forecast error



Notes: Shaded regions show the 10th-90th percentiles of absolute forecast errors.

Figure 2: Rounding behavior of forecasters



Notes:

with one decimal place. Given that there is much agreement between forecasters about the unemployment rate, a difference of 0.1 percentage points between two forecasts can be substantial. As such, we will ensure that the econometric model respects this rounding behavior, a form of non-classical measurement error.

3 Model

I assume that forecasters have access to a normally-distributed signal with some bias. That is, forcaster i's signal in period t is:

$$s^*_{i,t} \sim N(b_i + u_t, \nu^2_{i,t}) \tag{1}$$

where b_i is the forecaster's bias, u_t is the unemployment rate that they are attempting to forecast, and $\nu_{i,t}$ is the signal imprecision. We do not observe $s^*_{i,t}$, but instead observe the forecaster's *coarsening* of this signal $s_{i,t}$ (i.e., $s_{i,t}$ are the forecasts reported

in the survey). In particular, I assume that the forecaster reports $s_{i,t}$ rounded to an observable number of decimal points. Specifically:

1. If $s_{i,t}$ is reported as an integer, then $s^*_{i,t} \in (s_{i,t} - 0.05, s_{i,t} - 0.05)$, or
2. If $s_{i,t}$ is reported to d decimal places, then $s^*_{i,t} \in (s_{i,t} - 0.5 \times 10^{-d}, s_{i,t} + 0.5 \times 10^{-d})$.

I decompose the signal imprecision $\nu_{i,t}$ into a forecaster-specific component v_i and a time-specific component y_t, so that:

$$\nu_{i,t} = v_i \exp(y_t) \tag{2}$$

Here we can interpret v_i as the forecaster's own signal imprecision relative to other forecasters, and we can interpret y_t as measuring the difficulty of making a forecast in period t. I assume that y_t follows an AR(1) process:

$$y_t = \alpha y_{t-1} + \epsilon_t, \quad \epsilon_t \sim iidN(0, 1^2) \tag{3}$$

The normalization that $E[y_t] = 0$ and $V[\epsilon_t] = 1$ are needed becuase this is an unobserved, latent process. To see this, note that the model would make the same predictions if we (say) added one to all y_ts and then divided all v_is by e. One can interpret y_t as the normalized difficulty of making a forecast in period t.

Note that each forecaster has two parameters that describe their forecasting: bias b_i and imprecision v_i. While it is in principle possible to allow for these to be fully free parameters in the model, I instead choose to model them with a hierarchical specification. Specifically:

$$\begin{pmatrix} b_i \\ \log v_i \end{pmatrix} \sim \text{Multivariate Normal}(\mu, \Sigma) \tag{4}$$

Of particular interest in this model are estimated measures of forecast accuracy. For this, I shall use the estimated root-mean-squared prediction error based on parameters b_i and v_i, and assuming the long-run average level of forecast difficulty (i.e. $y_t = 0$):

$$\text{RMSPE}_i = \sqrt{b_i^2 + v_i^2} \tag{5}$$

I recover the parameters of the model using Bayesian techniques implemented in *Stan* (Carpenter et al., 2017).[1] Bayesian techniques have a few advantages over other

[1]Running sixteen chains, each with 50,000 iterations (25,000 after warm-up) resulted in diagnostics indicating that the posterior simulation had converged well enough. Specifically, all Gelman-Rubin statistics (Gelman and Rubin, 1992) where close to one and the effective sample size was sufficiently large for all parameters. Hamiltonian Monte Carlo diagnostics also revealed no pathological behavior of the chains (i.e. no divergent transitions after warm-up, no low BFMI chains).

approaches in this application. First, the latent processes for forecast difficulty and forecast rounding, and the forcaster-specific parameters b_i and v_i are especially well handeled by data-augmentation techniques. Second, once a posterior distribution of the model's fundamental parameters is simulated, computing the full posterior distributions of transformations like Equation 5 is straightforward.

4 Results

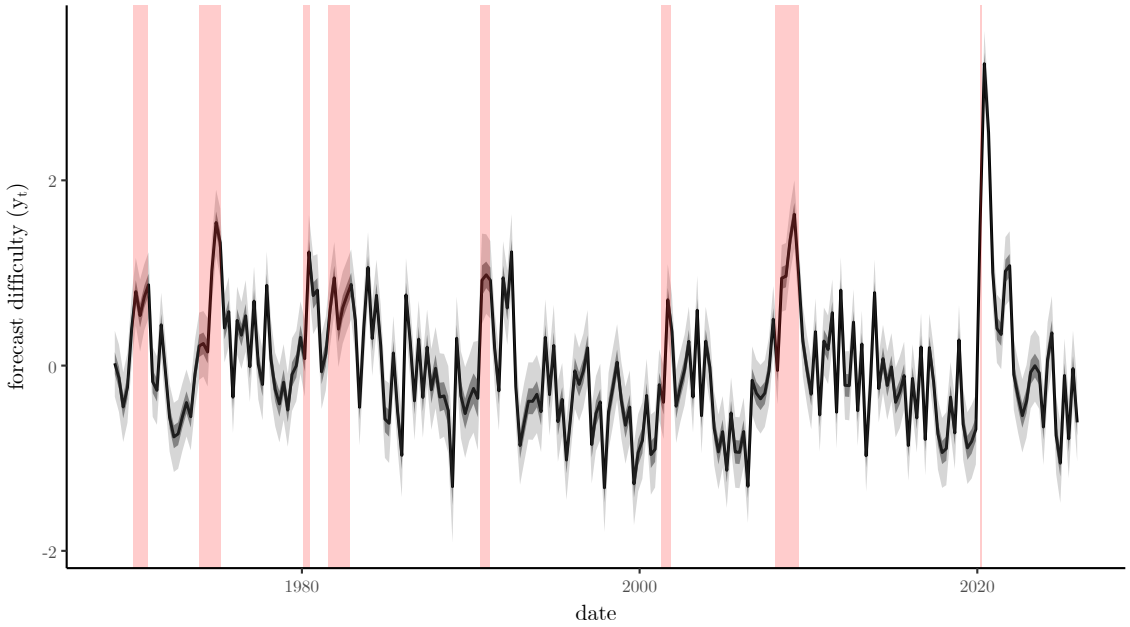
First, I present the latent measure of forecast difficulty. This is the variable that will equalize the playing field across time so that forecasters who happened to be in the sample when it was comparatively more difficult to forecast unemployment are not disadvantaged. Figure 3 plots the posterior mean of this latent time series (black line) alongside credible regions around this (gray shaded regions). Here we can see that the most difficult time to forecast unemployment was during the COVID19 pandemic, with other smaller peaks in difficulty showing up during recessions (shaded in red).

Figure 4 shows posterior estimates of individual forecasters' root-mean-squared prediction error (Equation 5). Forecasters are ranked from most precise (low rank) to least precise (high rank), based on their posterior mean estimates. While there is some heterogeneity in these variables, much of this heterogeneity is swamped by the posterior uncertainty (as shown in the error bars). *Marginal* posterior uncertainty, however, could be masking correlated posterior draws.[2] To explore this possibility, Table 1 shows the best ten forecasters ranked by their posterior mean root-mean-squared prediction error. The rightmost column of this Table shows the posterior probability that a forecaster is ranked first. Here we can see that about 49% of this probability mass is assigned to the first three forecasters, and 68% of this probability mass is assigned to the ten forecasters in this Table. Therefore we can be a little bit sure that the real top-ranked forecaster appears in this Table.

Figures 5 and 6 show the absolute forecast errors of the best three and worst three forecasters, respectively (colored lines). These are shown alongside the root-mean-squared forecast error for the entire sample (black line), and an 80% band for the absolute forecast error (gray region, 10th-90th percentiles). Note that the three best forecasters (Figure 5) are sperad out over different time periods. In fact, the third-ranked forecaster (yellow line) made forecasts during the most difficult forecasting

[2]Suppose for example that two random variables X_t and Y_t are related as follows: $Y_t = X_t + 1$. If the variance of these variables is large, then *marginally* it may look like there is a lot of posterior uncertainty in both Y_t and X_t, even though the model will be *very* sure that $Y_t > X_t$.

Figure 3: Forecast difficulty



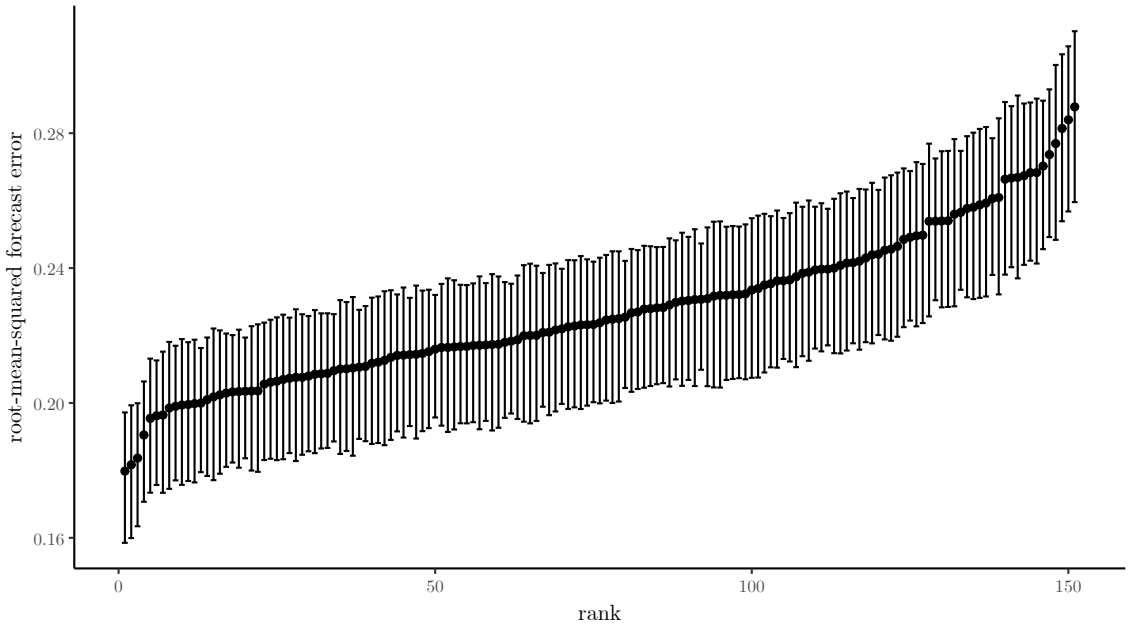
Notes: Black line shows the posterior mean estimate. Gray shaded regions are 50% and 95% Bayesian credible regions. Red shaded regions show recessions.

Table 1: Top ten forecasters

ID	time active	forecasts	RMSPE (model)	posterior probability ranked first
64	1968-12 - 1981-06	42	0.180	0.208
144	1970-12 - 1981-06	36	0.182	0.188
535	2005-06 - 2025-12	77	0.184	0.090
84	1968-12 - 2009-12	122	0.190	0.017
542	2005-06 - 2019-03	47	0.195	0.025
456	1994-12 - 2023-09	88	0.196	0.009
102	1969-09 - 1986-06	32	0.196	0.042
120	1970-12 - 1979-12	22	0.198	0.044
49	1968-12 - 1981-06	47	0.199	0.017
133	1971-03 - 1978-12	23	0.199	0.037

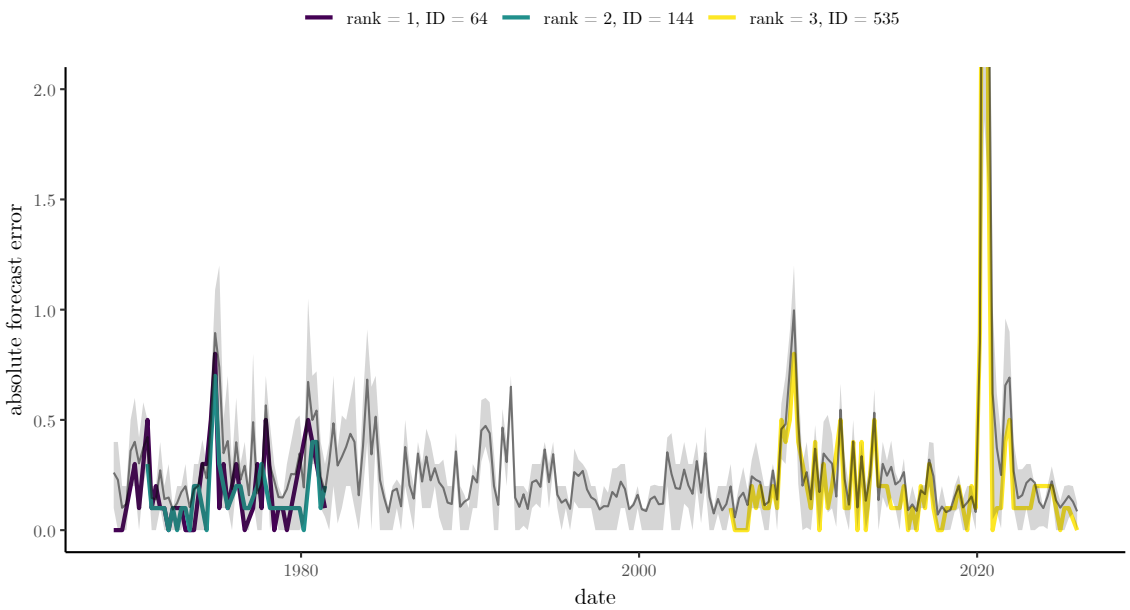
Notes:

Figure 4: Root-mean-squared forecast error



Notes: Dots show posterior mean estimates. Error bars show 50% Bayesian credible regions (25th-75th percentiles).

Figure 5: Best three forecasters



Notes: Black line shows root-mean-squared forecast error for the entire sample. Gray shaded region shows 10th-90th percentile of forecast error.

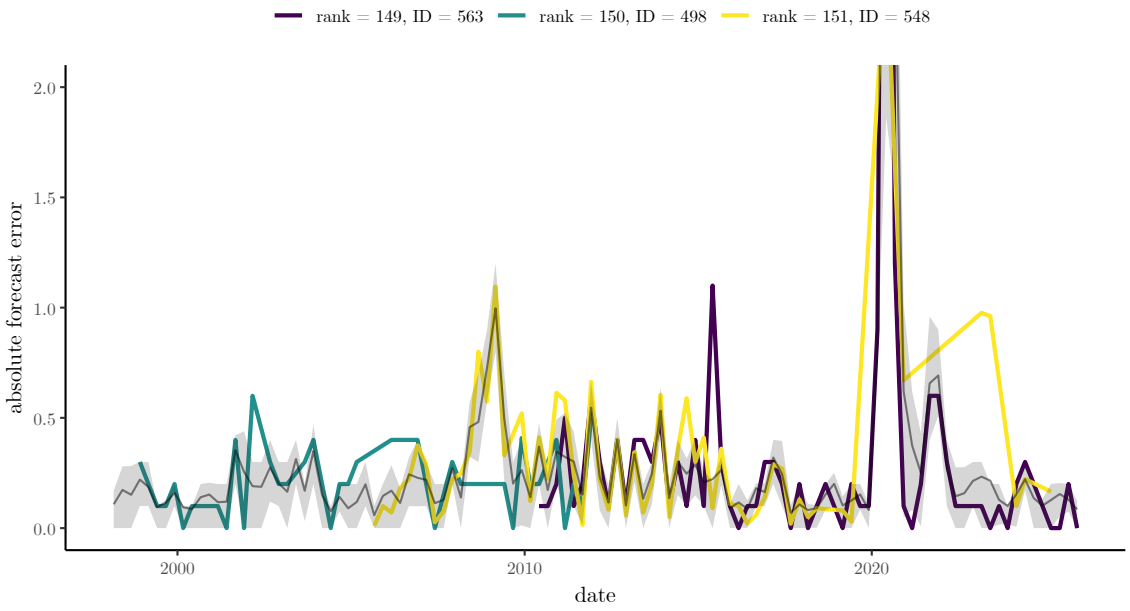
time, the COVID19 pandemic.

Finally, to show that the ranking is not only a function of observed forecast errors, Figure 7 compares the ranking based on the model (horizontal coordinate) to a ranking based on raw mean-squared prediction errors. There is no relationship between these two variables, suggesting that accounting for time-varying forecast difficulty is important in assessing forecasters' abilities.

5 Conclusion

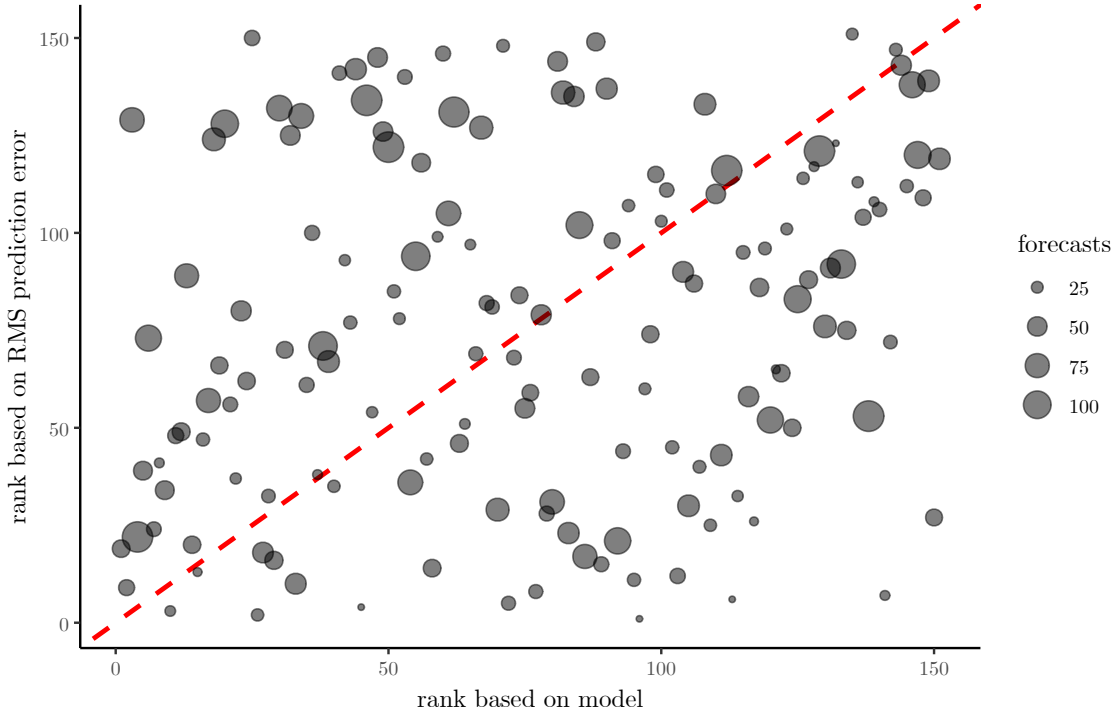
This paper proposes a method for assessing forecast accuracy in the face of time-varying forecast difficulty. In this situation raw forecast errors may be misleading, as some forecasters may make their forecasts in more difficult-to-forecast times than others. I model the evolution of forecast difficulty using a latent autoregressive process. The estimated forecast difficulty is estimated to have been greatest during the COVID19 pandemic, and greater during recessions than other times.

Figure 6: Worst three forecasters



Notes: Black line shows root-mean-squared forecast error for the entire sample. Gray shaded region shows 10th-90th percentile of forecast error.

Figure 7: Comparison of raw ranking and model ranking



Notes: Red dashed line is a 45° line.

While the model produces a difficulty-adjusted ranking of one hundred and fifty-one forecasters, the posterior estimates of individual accuracies are noisy. This suggests that either there is insufficient data to pin down these quantities, or that foreacsters are similarly accurate.

References

Bureau of Labor Statistics, U. (2026). Unemployment rate (UNRATE).

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of statistical software 76*.

Federal Reserve Bank of Philadelphia (2026). Survey of professional forecasters.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.