Assessing the Reliability of Probabilistic US Presidential Election Forecasts May Take Decades

Justin Grimmer[a,1], Dean Knox[b], and Sean J. Westwood[c]

[a]Democracy and Polarization Lab and Hoover Institution, Stanford University 616 Jane Stanford Way, Stanford CA,94305
[b]Analytics at Wharton and Department of Operations, Information and Decisions, University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104
[c]Department of Government, Dartmouth College, 3 Tuck Mall, Hanover, NH 03755
[1]Corresponding author: jgrimmer@stanford.edu

August 26, 2024

Abstract

Probabilistic election forecasts dominate public debate, drive obsessive media discussion, and influence campaign strategy. But in recent presidential elections, apparent predictive failures and growing evidence of harm have led to increasing criticism of forecasts and horse-race campaign coverage. Regardless of their underlying ability to predict the future, we show that society simply lacks sufficient data to evaluate forecasts empirically. Presidential elections are rare events, meaning there is little evidence to support claims of forecasting prowess. Moreover, we show that the seemingly large number of state-level results provide little additional leverage for assessment, because determining winners requires the weighted aggregation of individual state winners and because of substantial within-year correlation. We demonstrate that scientists and voters are decades to millennia away from assessing whether probabilistic forecasting provides reliable insights into election outcomes. Forecasters' claims of superior performance and scientific rigor should be tempered to match the limited available empirical evidence.

1 Introduction

Forecasts of U.S. presidential elections, despite recent failures, captivate the public and drive media narratives (Victor, 2020; Tufekci, 2020). The broad and intuitive appeal of presidential election forecasts, however, glosses over significant problems. A growing literature shows these forecasts can undermine the health of American democracy by facilitating consumer-driven horse-race election coverage (Victor, 2020; Mutz, 1995; Iyengar, Norpoth and Hahn, 2004), depressing turnout, and misleading the public about the viability of candidates (Westwood, Messing and Lelkes, 2020). The precise numeric claims made by forecasting practitioners, together with vocal claims of forecasting prowess, offer a veneer of scientific legitimacy (Victor, 2020). Yet quantitative forecasts have well-documented issues of their own and are not guaranteed to be accurate. Some issues undermine the polls that constitute these forecasts, like nonrepresentative samples (Williams and Reade, 2016), selective nonresponse because of campaign events (Gelman et al., 2016), variable (and unknown) polling accuracy across survey firms (Bon, Ballard and Baffour, 2019; De Stefano, Pauli and Torelli, 2022), failure to reach some members of the electorate (Erikson and Wlezien, 2008; Kennedy et al., 2018), or incorrect weights applied to responses (Kennedy et al., 2018). But even if the polls are correct, there can be drift in the relationship between polls and election outcomes over time (Munger, 2019; Grimmer, Roberts and Stewart, 2022) and there can be misspecification in the statistical model used to make predictions (Murphy, 2022). Without strong evidence of performance to support these claims, it is difficult to justify the public's fixation on forecasts.

Forecasters argue that quantitative models avoid certain common errors in conventional qualitative commentary, or punditry (Gelman et al., 2020). In criticizing conventional commentary, Silver (2012b) writes "[w]e need to stop, and admit it: we have a prediction problem. We love to predict things—and we aren't very good at it." Instead, Silver offers an alternative: probabilistic forecasting of election results based on recent polls, which is claimed to be superior (Silver, 2012*b*).

We demonstrate that currently available data are insufficient to empirically support probabilistic forecasters claims of being "good at it." This is because measures of a forecasting method's accuracy are themselves random variables. To infer that one method is superior to another, analysts require enough data to ensure the appearance of higher accuracy is not simply due to lucky guessing of election outcomes. Similar forecast models in finance, computer science, and machine learning are routinely evaluated with billions of training and validation observations (Card et al., 2020). In contrast, forecasting models in U.S. presidential elections necessarily rely on a mere handful of election results, because such elections occur only every four years.

Even though forecasters predict the outcomes of many states, those states contribute few effective observations, as we show in Section 3.2. One reason is that the forecasts across states are correlated, meaning that the results of elections across many states provides only limited information for learning the accuracy of election forecasts. A second reason this is true is that overall presidential election forecasts are weighted combinations of the underlying state forecasts. In fact, forecasts can perform better on average state-level predictions but worse in aggregated national-level prediction; this paradox arises because performance matters far more for swing states with many electoral college votes than for relatively easy-to-call or smaller states.

Claims of prowess and methodological rigor in election forecasts predictions are, as a result, necessarily not based on reliable data-driven evaluation. Of course, we can evaluate election forecasts based on other dimensions, such as their use of theoretically relevant predictive factors, the statistical soundness of their modeling strategy, or other features that might correlate with forecast accuracy. These dimensions can be useful for eliminating implausible forecasts (Lewis-Beck, 2005; Campbell, 2014). Ultimately, though, we cannot rule out plausible forecasts based solely on theoretical arguments: while one analyst might dismiss another's reasoning as ad hoc, the subjective nature of theoretical disputes often makes them difficult to adjudicate without empirical evaluation.

In best-case scenarios, we show that scientists are respectively decades, centuries, and millennia away from knowing if (1) probabilistic forecasting is more accurate than uninformed pundits guessing at random; (2) how well forecasters predict state and national election results; and (3) which techniques provide more accurate and well-calibrated predictions. If we were willing to make strong assumptions, we might be able to extrapolate from a forecaster's ability in other areas—say, sports or Congressional races—to gauge their ability in presidential elections. But this extrapolation is unlikely to be useful in practice, because presidential elections pose distinct challenges, and models that perform well in other settings do not necessarily perform well in presidential elections.

This lack of information also matters for resolving debates between two competing schools of probabilistic forecasters. We demonstrate that there has been no clearly superior method in recent presidential elections. Proponents of poll-based probabilistic forecasts (i.e., predictions that aggregate survey results to make predictions, often as late as the day before an election) have vocally claimed superiority over fundamentals-based forecasts (i.e., predictions that use information on the economy and incumbent performance to make a forecast, usually weeks to months before the election). But these claims have little basis in evidence: on average, fundamentals-based forecasts outperformed poll-based forecasts in 2016 and 2020. Of course, poll-based forecasts may ultimately prove more accurate. Our results show only that the jury is still out, and many more election results are needed to determine whether the recent empirical inferiority of poll-based forecasts is simply due to bad luck.

All in all, we demonstrate that statistical power matters when evaluating claims about forecasting prowess. Merely showing that one forecast method outperforms another on some metric in some year is not enough to conclude that the method has better performance in general. Rather, scholars must take care to present evidence that the differences in election forecasts are not solely due to luck or selective reporting.

2 How Do We Evaluate Forecasting Methods?

Political scientists have long predicted voting behavior using "fundamentals," such as economic growth, which empirically correlated with incumbents' margins of victory or defeat (Abramowitz, 1988; Lewis-Beck, 2005; Campbell, 2014). The goal in these fundamentals-based forecasts is to use information available weeks or months before an election to generate a prediction about who is likely to win.

More recently, Silver (2012*b*) transformed forecasting from an academic exercise into an industry that drives election coverage and shapes campaign narratives. Instead of using only fundamentals to forecast elections far in advance, a new group of forecasters—including *The Economist*, *The New York Times*, and the Princeton Election Consortium—use high-frequency polling data to make frequently updated and numerically precise statements about the probability of election outcomes, with election-eve predictions combining information about the fundamentals with the best polling data. These poll-based forecasters predict the winner of each state, each candidate's national vote share, candidates' electoral college votes, and ultimately each candidate's probability of victory.

Both poll- and fundamentals-based forecasters agree that empirical performance is the most important metric for evaluating these forecasts. Campbell (2008) asserts that "[t]he ultimate standard for any forecast or any forecasting model must be its accuracy," and in an essay evaluating fundamentals-based forecasts, Silver (2012a) focuses on two measures of their performance: mean absolute error and root mean square error. This notion of evaluating predictions against the predicted outcomes is not unique to elections. Such performance metrics are the standard measure of quality for predictive methods in general (Gneiting, 2011).

But elections have a particular problem: there are far too few events, and thus far too few predictions, for reliable assessment. Fundamentals-based forecasters have implicitly acknowledged this challenge by using additional, subjective criteria to compare similarly performing forecasts. For example, Lewis-Beck (2005) creates an index that combines accuracy, model

parsimony, reproducibility, and the number of days before an election that a forecast is made. Campbell (2008) argues this index is arbitrary, while Campbell (2014) argues for a related set of criteria. But these criteria generally only eliminate the most implausible methods; for example, these factors help rule out psychics and prescient barn animals, which lack transparency or reproducibility.

Subjective criteria and theoretical arguments are regarded as less useful for adjudicating between serious forecasting methods. Linzer (2014) argues, "The greatest impediment to the development of better election forecasting models is not a lack of theory; it is a lack of data." To address this gap, Linzer suggests collecting additional information about voters and states. But the underlying problem is that there are simply too few election outcomes. Even if we could measure thousands or millions of voter attributes, enormous differences between election cycles still remain; we could not possibly learn how to relate these attributes to vote choices in a general way. And while every presidential election produces results from 50 states and the District of Columbia, these elections are less useful due to unknown correlations between states and the fact that forecasters are ultimately interested in predicting an aggregate outcome.

Looming over forecasters, then, is a simple problem: the lack of evidence to support claims about forecasting prowess. We now characterize the severity of this problem by investigating, under best-case assumptions, the number of additional elections needed before commonly made claims would be defensible at standard levels of statistical significance.

3 Assessing U.S. Election Forecasts

How many presidential elections would analysts need to observe before drawing firm conclusions about the performance of forecasting methods? We now present analytic results and naturalistic simulations that demonstrate that the required volume of evaluation data outstrips the currently available amount by orders of magnitude. In Section 3.1, we present

simple analyses showing that society is two decades to a century away from assessing whether presidential election forecasters are more accurate than an uninformed, randomly guessing pundit at conventional levels of statistical significance. This remains true even when we focus on state-level predictions from forecasters—though we caution that better performance on state-level forecasts does not necessarily imply better performance on aggregated predictions. To demonstrate this, in Section 3.2, we turn to rich, naturalistic simulations of election outcomes. We further show that even under highly implausible best-case assumptions, currently competing forecasters will not be statistically distinguishable on standard performance metrics for decades or millennia to come. We then analyze the performance of actual forecasters and show that neither poll-based nor fundamental-based forecasts have produced dominant performance. Based on these results, we conclude that vigorous ongoing disputes over scientific rigor between various forecasters are not grounded in empirical evidence of the forecasters' performance.

3.1 Determining if a Forecaster is Superior to Random Guessing

Forecasters' primary goal is to predict the winner of a presidential election. To assess the informativeness of succeeding or failing at this task, we draw on Silver's (2012*b*) characterization of uninformed pundits as being as good as a "coin flip." Specifically, we ask: how many elections would it take before a skilled forecaster consistently outperformed a pundit guessing uniformly at random, in terms of historical track record? For transparency, we make the maximally generous assumptions that (1) each forecaster has fixed skill, or probability of being correct, and therefore accumulates successes according to a binomial distribution based on the number of elections and their skill; (2) that the pundit has no access to the forecaster's information, so the predictions are independent; and (3) outcomes of successive elections are independent. Together, these assumptions ensure that each cycle provides the largest amount of information possible. We then vary the "skill" of the forecaster from 55% accurate (5 percentage points better than the pundit) to 95% (45 percentage points better).

Simulation details are given in Appendix A.1.

Figure 1 shows the probability that the forecaster has called more elections correctly than the coin-flip pundit over the first ten elections, along with the number of elections needed before the forecaster appears superior with 95% probability. Depending on forecaster skill, distinguishing their performance from random guessing at generally accepted levels of statistical significance will take many years. For example, suppose forecasters are as accurate as the FiveThirtyEight forecast over the past four cycles ($\frac{3}{4}$ correct, or 25 percentage points better than coin-flip pundits). After one cycle (four years), the forecaster is winning with 37.5% probability—Pr(forecaster correct) $\times$ Pr(pundit incorrect)—and losing with 12.5% probability. The forecaster only exceeds a 95% win probability threshold after 24 cycles (96 years). For a 55%-accurate forecaster, 559 elections (2,236 years) are needed for its record to surpass the coin-flip pundit with 95% probability. (If we imagine infinite coin-flipping pundits rather than infinitely repeated trials, then these reported "win probabilities" are the percentage of coin-flipping pundits that the forecaster outperforms.)

Next, we extend our results beyond raw accuracy to assess whether forecasters can meaningfully compete on *calibration*—that is, whether the events in fact occur as frequently as they are predicted to occur. Again, we will proceed by examining an unrealistically generous case. Suppose (1) the true probability of a victory is a highly informative 0.89, corresponding to the final FiveThirtyEight forecast for a Biden victory in 2020; (2) independent and identically distributed elections recur cycle after cycle; and (3) our protagonist forecaster is "oracular" in the sense of being perfectly calibrated. We contrast this oracular forecaster with the coin-flip pundit and other competing forecasters who have higher, but still inferior, skill. Under these parameters we simulate a large number of election results and assess performance by computing each forecasters' (1) average absolute error and (2) average squared error, or Brier score.[1] Simulation details are given in Appendix A.2.

Even in these ideal circumstances, Table 1 shows that seven presidential elections (28

[1]Repeatedly "rerunning" an election in this way is analogous to asking: holding all else equal, how many forecasts would we have to observe to determine a forecaster is better calibrated at a particular level?

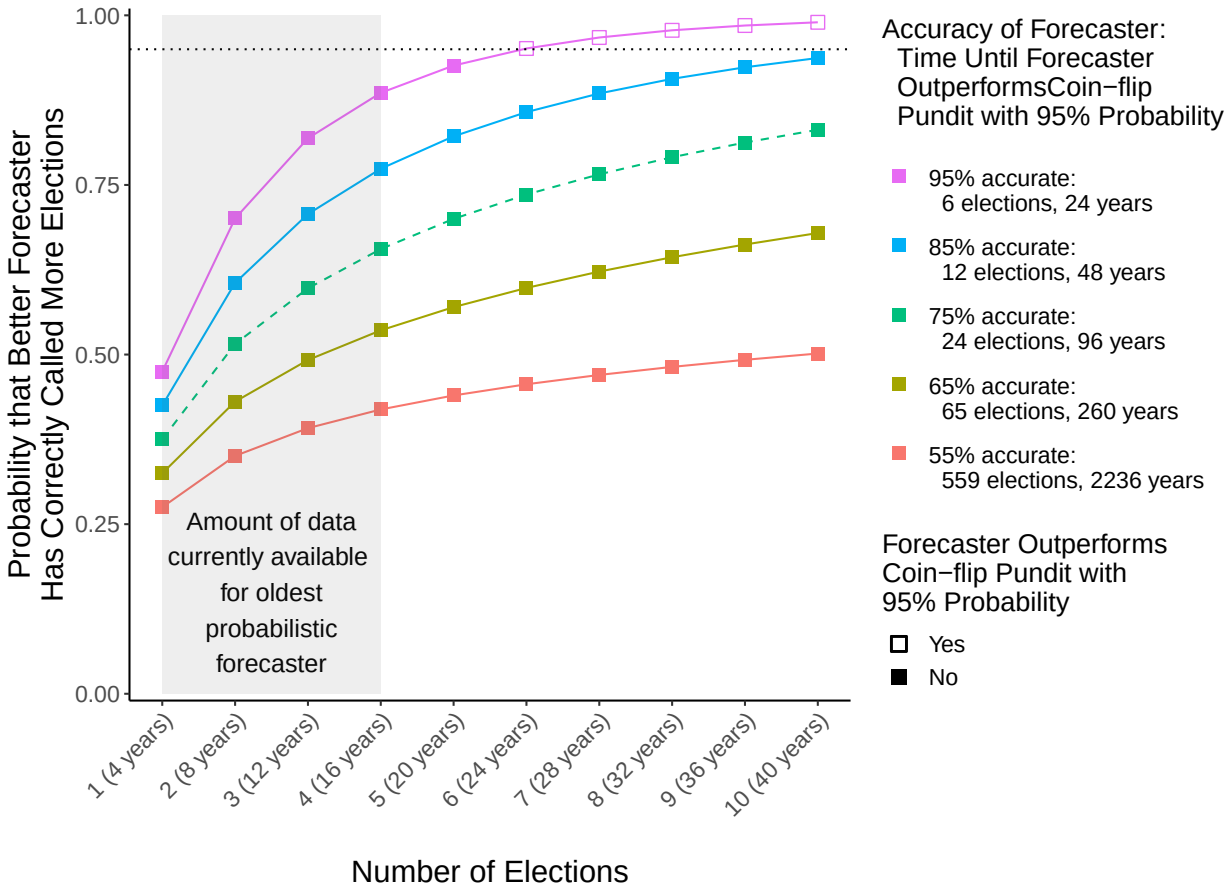

Figure 1: **Accuracy comparison of skilled forecaster to coin-flip pundit.** Probability that a forecaster achieves a superior empirical record than a randomly guessing pundit, for various levels of forecaster skill and number of recorded election cycles. The forecaster with 75% accuracy (empirical FiveThirtyEight record) is indicated with a dashed line.

years) are required before the oracle consistently outperforms the coin-flip pundit (has >95% probability of attaining a superior empirical record), regardless of whether absolute error or Brier score is used to assess calibration performance. More nuanced comparisons between less naïve competitors, however, require a massive set of elections. When compared to an 85%-accurate alternative, the 89%-accurate oracle only attains a consistently superior track record after 2,588 years.

Competitor Skill	Years to Separation
0.50	28
0.55	32
0.60	32
0.65	72
0.70	100
0.75	204
0.80	492
0.85	2,588

Table 1: **Comparing calibration of oracular forecaster to competitors of varying skill.** We compare an oracular forecaster, achieving the best possible skill of 89% accuracy given inherent noise in election outcomes, against less skilled competitors. The right column reports the number of years before the oracle has >95% chance of attaining a superior record. Results are identical regardless of whether absolute-error or Brier-score evaluations are used.

3.2 Determining If One Forecaster Is Superior to Another

U.S. Presidential elections are a composite of 56 state- or congressional-district-level election results (51 state-level results in each state and the District of Columbia, with additional district-level results in Maine and Nebraska that influence their vote allocation). These lower-level election winners are aggregated by a weighted sum that determines the overall national-level winner, via the U.S. Electoral College. Moreover, within each state, presidential votes are counted within counties and precincts. It might appear that the results of individual elections provides an opportunity to have many tests of the accuracy or calibration of election forecasting methods. Yet, the results across or within states are often correlated, particularly given swings in the election results. For example, in the 2016 election Donald Trump won

several swing states that were a very small margin. But in 2020, Joe Biden won many of those same states by a similarly small margin. As a result of the correlation across units, there are many fewer effective observations than the total number of states, counties, or precincts used to assess the accuracy of the method.

We begin by using a simplified analysis to build intuition for why the large number of lower-level elections does not translate into a precise estimate of national-level forecast accuracy. For simplicity, we start by making the implausibly generous assumption that a forecasting method has the same probability of correctly predicting the winner of every lower-level election, so that each lower-level election provides information about the method's underlying accuracy. Suppose in state i in election t, the forecasting method either correctly selects the winner ($X_{i,t} = 1$) or makes an error ($X_{i,t} = 0$). The true, but unknown, accuracy of the forecaster is θ for all elections, $\mathbb{E}[X_{i,t}] = \theta$ for all i. The results across states, counties, and precincts are almost certainly correlated within a particular election t—for example, due to systematic nonresponse to polling, cross-location turnout efforts from campaigns, or misspecification of the forecasting model. Thus, whether the method accurately predicts the result of the election is also likely to be correlated across the lower-level units. For conveying intuition in this simple example, we will further suppose that the method's forecast performance has the same dependence across every pair of units within the same election t, so that the $\text{Corr}(X_{i,t}, X_{j,t}) = \rho$ for all $i \neq j$ in the same election t, and $\text{Corr}(X_{i,t}, X_{j,t'}) = 0$ for all $t \neq t'$. We also suppose there is a constant variance in our prediction's classification errors, $\sigma^2_{i,t} = \sigma^2$ for all states i and elections t. After observing T elections in N units, the proportion of correctly forecasted election results is $\widehat{\text{Accuracy}} = \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{X_{i,t}}{N \times T}$. It can be shown that the variance of $\widehat{\text{Accuracy}}$ is

$$\text{Var}(\widehat{\text{Accuracy}}) \quad = \quad \frac{\sigma^2(1-\rho)}{N \times T} + \frac{\sigma^2 \rho}{T} \tag{1}$$

Equation 1 shows that reducing $\text{Var}(\widehat{\text{Accuracy}})$ requires more than just increasing the number of units N if there is a positive correlation in the prediction errors across units. For a fixed correlation ρ, the first term in Equation 1 grows smaller—meaning a forecaster's accuracy is more precisely estimated—if there are more units N or more elections T. However, no matter how many state-, district-, or county-level contests are predicted within a single election, the second term cannot be resolved, because the correlation in prediction errors across these lower-level results implies that many elections need to be observed to precisely estimate the forecaster's accuracy.

This analysis is a simplification of the actual problem. In real-world elections, we have much less information about accuracy than Equation 1 implies. This is because forecasting methods do not have constant accuracy across U.S. states. Some states are straightforward to predict given the comfortable margin for either party in that state, while other states are tossups that are difficult to predict. Second, moving to lower units—such as counties or precincts—will almost certainly increase the correlation in the accuracy of predictions across units. Very similar idiosyncratic polling errors, model misspecifications, or unobserved campaign efforts are likely to be found across the lower level units. And as a result, Equation 1 shows that adding these additional units will provide limited additional information about a forecaster's accuracy. Third, forecasting the winner of the overall election requires aggregating state-level predictions to determine a final winner. As we show in the next section, more accuracy at the state-level does not necessarily imply greater accuracy in predicting the overall winner.

3.2.1 Naturalistic Simulation To Determine Forecasting Superiority

We now present a naturalistic simulation of actual electoral outcomes to assess whether probabilistic forecasters can be meaningfully assessed using finer-grained, state-level predictions. As before, we consider an unrealistically generous setting. Specifically, we use the

estimated FiveThirtyEight 2020 election model as our true data-generating process.[2] Sequences of election outcomes are drawn from this population, one map at a time, to produce simulated election chains extending 4,000 years into the future. Thus, our simulation includes correlations in state-level outcomes. Importantly, the election model is assumed to remain constant over time, eliminating what is perhaps the largest source of error in real-world forecasting. The correlations across state limits the information we can learn about a forecasters' performance in each state, because the correlation in state results implies a forecasters' performance across states will be correlated.

Using this simulated world, we examine five competing election forecasts. Our protagonist is (1) the "oracle" FiveThirtyEight 2020 forecast on which the simulation is based. This oracle has perfect information about the *expected* outcomes of state and national elections; its only source of error is the remaining inherent randomness, or unknowable variation, in the process. We compare this oracular forecaster to (2) The Cycle (Bitecofer, 2020), (3) The Economist (Morris and Gelman, 2020), (4) PredictIt (Staff, 2020), and (5) FiveThirtyEight's outdated 2016 forecast (Silver, 2016). These competing forecasts, by design, will underperform the oracular forecast because the simulation presumes that the oracle model has perfect knowledge about the non-random elements of the simulated world. Thus, our results do not indicate the ground-truth quality of each model; rather, we use these simulations to probe whether competing forecasters can be reliably distinguished from one another.

Our simulation will reflect an idealized scenario in which there is no temporal drift in the data-generating process and the "all else equal" condition is satisfied across elections. In other words, we will make a series of best-case assumptions in order to isolate the phenomenon of interest. Specifically, we will (unrealistically) assume that fundamentals and polling results unfold in exactly the same way across election years; as a result, each forecaster will draw the same conclusion in every cycle. The only remaining source of variation in our results is the inherent randomness in the election outcome. Simulation details are

[2]Our population of electoral outcomes consists of simulated full maps from the FiveThirtyEight model (Gelman, 2020).

given in Appendix A.3.

We evaluate forecasts using four criteria: mean absolute error and mean squared error of all historical electoral college vote predictions, and accuracy and calibration of all historical state-level predictions (Brier score). For each metric, "running tallies" average over all prior years and (where relevant) states.[3]

Figure 2 visualizes the Brier score of four competing forecasters, relative to the oracle. Colored horizontal lines indicate the expected performance difference at each point in time; a 95% envelope, based on Monte-Carlo simulations, indicates the distribution of possible differences in forecasting records between the competitor and the oracle. When this envelope no longer overlaps zero (dashed black line), analysts can expect to reject the null hypothesis (that a competitor is no better or worse than the oracle) at conventional levels.

We find that several decades of election results are needed to reliably distinguish even an oracular forecaster from weaker (by design) competitors. When evaluated against FiveThirtyEight's 2020 model, The Cycle's forecast requires 17 elections (68 years) to separate; The Economist's forecast requires 21 elections (84 years); and PredictIt, 22 elections (88 years). Even a "forecast" that simply regurgitates FiveThirtyEight's outdated 2016 predictions cannot be reliably distinguished in less than 14 elections (56 years).

Table 2 reports the time to distinguishability for all forecasters and evaluation criteria. When we focus on an overall national metric, such as the historical average absolute error in Electoral College forecasts, we find that 2,612 years are required before a perfect oracular forecast will separate from The Cycle's forecast; 176 years are needed for PredictIt, and more than 4,000 years (the maximum duration examined) for the Economist. Even when examining historical state-level accuracy, centuries of elections are required. (We caution that state-level forecasts are useful assessments of a forecast, but higher forecaster accuracy at the state level does not necessarily imply higher accuracy when selecting a national winner.)

[3]For electoral-vote forecasts that do not disaggregate predictions by district, we treat Maine and Nebraska prediction as blockwise predictions for all in-state districts. For state-level performance metrics, sub-state districts are omitted for all forecasters.

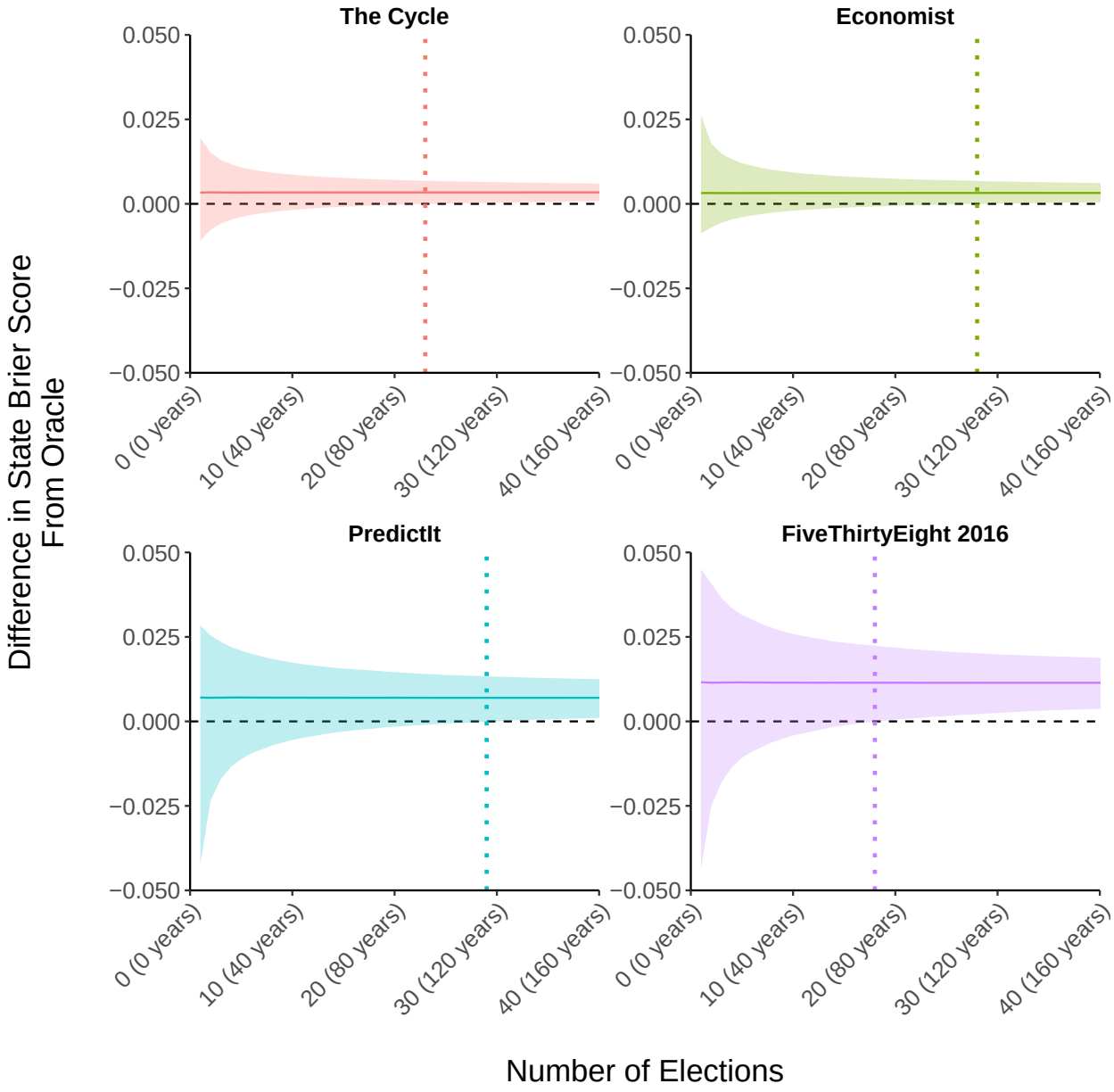

Figure 2: **Calibration comparison of competing forecasters, relative to an oracular forecaster.** Forecasters are evaluated in an election simulation that randomly generates all district outcomes from a naturalistic model. Each panel depicts a competing forecaster. The simulation repeats for sequences of many elections, progressing along the x-axis. The y-axis depicts the running difference in state-level Brier scores, one of the performance metrics, up until a particular point in time. Horizontal colored lines indicate running performance gaps between a competitor and the oracle. Each vertical slice of the shaded regions contains 95% of possible differences in historical track records that can emerge at a given point in time. Forecasters can be reliably distinguished when this 95% envelope no longer contains the black dashed line, indicating zero difference. The earliest distinguishable time is shown with a vertical dotted line.

Forecaster	EV Pred. (Abs. Error)	EV Pred. (Sq. Error)	State Pred. (Calibration)	State Pred. (Accuracy)
Cycle	653 elections	529 elections	17 elections	55 elections
	2,612 years	2116 years	68 years	220 years
Economist	>1,000 elections	302 elections	21 elections	317 elections
	>4,000 years	1,208 years	84 years	1,268 years
PredictIt	44 elections	33 elections	22 elections	23 elections
	176 years	132 years	88 years	92 years
538(2016)	39 elections	29 elections	14 elections	27 elections
	156 years	116 years	56 years	108 years

Table 2: **Time to distinguishability between oracular forecaster and competitors.** Forecasters are evaluated in a naturalistic election simulation of all district outcomes, repeating for sequences of 1,000 elections (4,000 years). Each row denotes a competing forecaster, with columns indicating the time required before the superior (by construction) performance of the oracle forecaster is distinguishable at the conventional 95% level. EV refers to Electoral College Votes.

We reiterate that this simulation is highly conservative and will necessarily understate the time needed to distinguish which forecast performs better. This is because in truth, the underlying characteristics of elections are guaranteed to change over elections. Our simulations abstract away this source of variation; instead, they seek to isolate the role of inherent stochasticity. If continual changes in underlying conditions mean that different forecasts can outperform competitors at different points in time, as seems likely, then it will take many more elections to distinguish them.

4 Comparing Fundamentals- and Poll-based Forecasts

Our results show that society lacks sufficient information to determine if one forecasting method is superior than another. Even so, if one mode of forecasting *appears* superior to another, some might prefer it. For example, less scientifically minded consumers might be willing to set aside concerns about statistical validity. Indeed, Silver (2012*a*) makes precisely this argument when claiming fundamentals-based forecasts are historically inaccurate with respect to two-party vote shares in presidential elections, implicitly arguing for poll-

based forecasts. Silver (2012*a*) further critiques the lack of consensus between different fundamentals-based forecasts, arguing that this constitutes evidence of their flaws. (Note, however that poll-based forecasts also exhibit substantial variability; this simply reflects data paucity and the arbitrariness of modeling assumptions.)

To examine this possibility, we now assess whether one approach might currently appear preferable. Table 3 offers an updated version of Silver's (2012*a*) comparison between (1) fundamentals-based forecasts and (2) various poll-based forecasts, using published results from the "Campbell collection" of presidential election forecasts (Cuzán, 2020). Our approach averages over all fundamentals-based forecasts, rather than examining each one in isolation. In each year, we compare forecasted two-party vote share to the true presidential election result. Poll-based numbers are based on the last available forecasts before the election.

	Fundamentals (Aggregated)		Poll-Based (Individual)		
Year	Number	Avg. Error	FiveThirtyEight	Princeton	Economist
1992	6	3.35			
1996	12	−1.39			
2000	9	5.01			
2004	15	2.87			
2008	16	2.07	−0.20	−0.30	
2012	12	−1.32	−0.70	−0.90	
2016	10	−0.59	0.80	0.90	
2020	10	0.09	−1.75	−0.35	−2.10

Table 3: **Comparing Fundamentals-Based Forecasts to Poll-Based Forecasts of Two-Party Vote Share.** In 2008 and 2012, poll-based forecasters achieved a low average absolute error (<0.1 percentage points for the two-party vote share on average), outperforming the average fundamentals-based forecast (1.7 percentage points). However, in 2016 and 2020, poll-based forecasters performed poorly (1.2 percentage points), markedly underperforming the average fundamentals-based forecast (0.3 percentage points). In fact, no individual poll-based forecaster can compete in these years. Over the full 2008–2020 period for which data is available, the average fundamentals- and poll-based forecasts achieved virtually identical absolute error rates (1.0 versus 0.9 percentage points).

In 2008 and 2012, poll-based forecasts were closer to the ultimate two-party vote share, relative to the average fundamentals-based forecast. But in both 2016 and 2020, poll-based

forecasts performed notably poorly; this remains true whether examining poll-based forecasters individually or in the aggregate. All in all, over the four elections with available data, the two approaches have performed virtually identically.

Thus, even setting statistical principles aside, the evidence does not suggest that the poll-based forecasts have clearly improved over more traditional fundamental based forecasts. This finding is particularly surprising in light of the fact that poll-based forecasts use vast amounts of information, including polls conducted up to the day before the election, whereas fundamentals-based forecasts use only a handful of data points. However, rigorously assessing which method provides better results will require many more elections.

5 Why are We Making Probabilistic Forecasts?

Even under the most optimistic assumptions, we are far from being able to rigorously assess probabilistic forecasters' claims of superiority to conventional punditry. Yet despite the lack of demonstrable benefits, forecasts induce known harms: producing vacuous, unverifiable horse-race coverage; potentially depressing votes for forecasted winners; and misleading the public and campaigns alike. Taken together, it is hard to justify the place of forecasts in the political discourse around elections without fundamentally recalibrating claims to match the available empirical evidence.

Recalibrating forecasts requires a more extensive accounting of the errors that go into predictions. An incomplete list of additional sources of variance in predictions would include (1) uncertainty over the correct model specification, including how to accurately account for covariance between states; (2) additional difficult-to-quantify analyst degrees of freedom; (3) the fact that prediction models necessarily over-smooth due to a lack of training data, for example by estimating separate state and voter-race coefficients rather than interacting these factors; (4) temporal drift in survey nonresponse patterns, voting-intent misreporting patterns, and vote-choice patterns; and (5) underlying sampling variability in representating

and nonrepresentative polling.

Currently, only a fraction of these sources of error are incorporated into uncertainty measures, making reported confidence intervals highly overoptimistic relative to long-run error rates. After addressing these issues, true uncertainty in election forecasts will necessarily be much wider than currently advertised. For forecasters, this may appear unsatisfying in the short run, as it inhibits the lucrative marketing of statistical expertise for predicting election outcomes. But an honest accounting of the limitations in forecasts is critical to long-run faith in the entire forecasting enterprise and, we argue, for helping voters understand the inherent lack of precision in the models that currently play an outsize role in American democracy.

References

Abramowitz, Alan I. 1988. "An improved model for predicting presidential election outcomes." *PS: Political Science & Politics* 21(4):843–847.

Bitecofer, Rachel. 2020. "15 Months In, The Negative Partisanship Model Predictions for Presidential Race/Congress Have Come To Pass.".

Bon, Joshua J, Timothy Ballard and Bernard Baffour. 2019. "Polling bias and undecided voter allocations: US presidential elections, 2004–2016." *Journal of the Royal Statistical Society Series A: Statistics in Society* 182(2):467–493.

Campbell, James E. 2008. "Evaluating US presidential election forecasts and forecasting equations." *International Journal of Forecasting* 24(2):259–271.

Campbell, James E. 2014. "Issues in presidential election forecasting: Election margins, incumbency, and model credibility." *PS: Political Science & Politics* 47(2):301–303.

Card, Dallas, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald and Dan Jurafsky. 2020. With Little Power Comes Great Responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online:

Association for Computational Linguistics pp. 9263–9274.
URL: *https://www.aclweb.org/anthology/2020.emnlp-main.745*

Cuzán, Alfred G. 2020. "The Campbell Collection of Presidential Election Forecasts, 1984–2016: A Review." *PS: Political Science & Politics* pp. 1–5.

De Stefano, Domenico, Francesco Pauli and Nicola Torelli. 2022. "Preelectoral polls variability: A hierarchical Bayesian model to assess the role of house effects with application to Italian elections." *The Annals of Applied Statistics* 16(1):460–476.

Erikson, Robert S and Christopher Wlezien. 2008. "Leading economic indicators, the polls, and the presidential vote." *PS: Political Science & Politics* 41(4):703–707.

Gelman, Andrew. 2020. "Reverse-engineering the problematic tail behavior of the Fivethirtyeight presidential election forecast.".

Gelman, Andrew, Jessica Hullman, Christopher Wlezien and George Elliott Morris. 2020. "Information, incentives, and goals in election forecasts." *Judgment and Decision Making* 15(5):863.

Gelman, Andrew, Sharad Goel, Douglas Rivers, David Rothschild et al. 2016. "The mythical swing voter." *Quarterly Journal of Political Science* 11(1):103–130.

Gneiting, Tilmann. 2011. "Making and evaluating point forecasts." *Journal of the American Statistical Association* 106(494):746–762.

Grimmer, Justin, Margaret E Roberts and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Iyengar, Shanto, Helmut Norpoth and Kyu S Hahn. 2004. "Consumer demand for election news: The horserace sells." *The Journal of Politics* 66(1):157–175.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers et al.

2018. "An evaluation of the 2016 election polls in the United States." *Public Opinion Quarterly* 82(1):1–33.

Lewis-Beck, Michael S. 2005. "Election forecasting: Principles and practice." *The British Journal of Politics and International Relations* 7(2):145–164.

Linzer, Drew A. 2014. "The future of election forecasting: More data, better technology." *PS, Political Science & Politics* 47(2):326.

Morris, G. Elliott and Andrew Gelman. 2020. "Forecasting the US elections.".

Munger, Kevin. 2019. "The limited value of non-replicable field experiments in contexts with low temporal validity." *Social Media+ Society* 5(3):2056305119859294.

Murphy, Kevin P. 2022. *Probabilistic machine learning: an introduction*. MIT press.

Mutz, Diana C. 1995. "Effects of horse-race coverage on campaign coffers: Strategic contributing in presidential primaries." *The Journal of Politics* 57(4):1015–1042.

Silver, Nate. 2012*a*. "Models Based on 'Fundamentals' Have Failed at Predicting Presidential Elections." *538.com* . https://fivethirtyeight.com/features/models-based-on-fundamentals-have-failed-at-predicting-presidential-elections/.

Silver, Nate. 2012*b*. *The signal and the noise: why so many predictions fail–but some don't*. Penguin.

Silver, Nate. 2016. "FiveThirtyEight 2016 Election Forecast.".

Staff. 2020. "Predictable Insights - 10.30.20.".

Tufekci, Zeynep. 2020. "Can We Finally Agree to Ignore Election Forecasts?".

Victor, Jennifer Nicoll. 2020. "Let's Be Honest about Election Forecasting." *PS: Political Science & Politics* pp. 1–3.

Westwood, Sean Jeremy, Solomon Messing and Yphtach Lelkes. 2020. "Projecting confidence: How the probabilistic horse race confuses and demobilizes the public." *The Journal of Politics* 82(4):1530–1544.

Williams, Leighton Vaughan and J James Reade. 2016. "Forecasting elections." *Journal of Forecasting* 35(4):308–328.

Appendix

Table of Contents

A Appendix

A.1 Section 3.1 Accuracy Simulation Details

Suppose forecaster i's accuracy, or skill, is π_i. Define the "hard," or binary, forecast in election t as $\hat{Y}_{i,t} \in \{0, 1\}$; it matches the true election result, $Y_t \in \{0, 1\}$, with probability $\Pr(\hat{Y}_{i,t} = Y_t) = \pi_i$. After T elections, the number of correctly called elections is $Z_{i,T} = \sum_{t=1}^{T}(\hat{Y}_{i,t} = Y_t)$, and the running proportion is $\hat{\pi}_{i,T} = \frac{Z_{i,T}}{T}$. We compare the performance of forecaster i to competitor i', a coin-flipping pundit with skill $\pi_{i'} = 0.5$, forecasts $\hat{Y}_{i',t}$ and track record $\hat{\pi}_{i',T}$. We are interested in determining the smallest T such that $\Pr(\hat{\pi}_{i,T} > \hat{\pi}_{i',T}) \geq 0.95$, or equivalently $\Pr(Z_{i,T} > Z_{i',T})$.

To do so, we analytically compute the probability that i is winning after T elections—i.e. holds a superior track record—then vary T. The number of successes for both the forecaster and the pundit follow binomial distributions, $Z_{i,T} \sim \mathcal{B}(T, \pi_i)$ and $Z_{i',T} \sim \mathcal{B}(T, \pi_{i'})$. The win probability is given by $\sum_{z=1}^{T}\sum_{z'=0}^{z-1}\Pr(Z_{i,T} = z)\Pr(Z_{i',T} = z') = \sum_{z=1}^{T}\sum_{z'=0}^{z-1}\binom{T}{z}\pi_i^z(1 - \pi_i)^{T-z}\binom{T}{z'}\pi_{i'}^{z'}(1 - \pi_{i'})^{T-z'}$.

A.2 Section 3.1 Calibration Simulation Details

We now examine "soft," or continuous forecasts with $\hat{Y}_{i,t} \in [0,1]$, then analyze the time needed to detect calibration differences. We hold "all else equal," so that forecaster i always predicts $\hat{Y}_{i,t} = \pi_i$. The true election result, Y_t, is assumed to follow a Bernoulli distribution with $\Pr(Y_t = 1) = 0.89$.

For each forecaster i, after T elections, we compute the running mean absolute error $\text{MAE}_{i,T} = \frac{1}{T}\sum_{t=1}^{T} |\hat{Y}_{i,t} - Y_t|$ and mean squared error $\text{MSE}_{i,T} = \frac{1}{T}\sum_{t=1}^{T} (\hat{Y}_{i,t} - Y_t)^2$. It can be shown that the "oracle" forecaster with $\pi_i = 0.89$ will achieve perfect calibration—i.e., $\pi_i = \mathbb{E}[Y_t]$—and thus achieve the lowest possible average error, whether absolute or squared. We compare the oracle forecast to other uncalibrated, inferior forecasts with varying skill parameters.

First, observe that for forecaster i, $\text{MAE}_{i,T}$ can be rewritten as $\text{MAE}_{i,T}(\pi_i, Z_T) = \frac{1}{T} Z_T(1-\pi_i) + \frac{1}{T}(T - Z_T)\pi_i$, where $Z_T = \sum_{t=1}^{T} Y_t$ is the number of elections with $Y_t = 1$. This follows $Z_T \sim \mathcal{B}(T, 0.89)$. The probability that oracular forecaster i achieves a better empirical MAE than competitor i' is

$$\begin{aligned}
&\Pr\Big(Z_T \in \{z : \text{MAE}_{i,T}(\pi_i, z) < \text{MAE}_{i',T}(\pi_{i'}, z)\}\Big) \\
&\qquad = \sum_{z=0}^{T} \binom{T}{z} 0.89^z \; 0.11^{T-z} \; \Big\{\text{MAE}_{i,T}(\pi_i, z) < \text{MAE}_{i',T}(\pi_{i'}, z)\Big\}.
\end{aligned}$$

We analytically compute this quantity for increasing T until the oracle achieves a win probability on mean absolute error that exceeds 95%. We repeat for the mean squared error (Brier score) metric and obtain identical results, because squared error is a monotonic transformation of absolute error.

A.3 Section 3.2 Simulation Details

Our naturalistic district-level analysis proceeds as follows. We will initially consider a single simulated world, suppressing the simulation index until later. Let $Y_{j,t} \in \{0,1\}$ denote the

outcome of district $j \in \{1, \ldots, J\}$ in election t. We collect these in $\boldsymbol{Y}_t = [Y_{1,t}, \ldots, Y_{J,t}]^\top$ and let $\boldsymbol{Y}_t \sim$ FiveThirtyEight; that is, in each election, we simulate entire maps from the fitted FiveThirtyEight forecast model for 2020. This data-generating process therefore allows for correlations between state outcomes. From the district-level outcomes, we also obtain EV_t, the cumulative electoral votes for the first candidate.

We will consider five competing forecasters. The oracle, i, simply predicts $\hat{\boldsymbol{Y}}_{i,t} = \mathbb{E}[\boldsymbol{Y}_t]$, i.e., the actual set of district-level predictions made by FiveThirtyEight in 2020; for the national outcome, it predicts $\hat{\text{EV}}_{i,t} = \mathbb{E}[\text{EV}_t]$. These predictions remain the same in every election, because we make the maximally generous assumption that each election unfolds identically in terms of fundamentals and polling results. Similarly, competing forecasts consist of the actual state and electoral vote predictions made by The Cycle, The Economist, PredictIt, and FiveThirtyEight's outdated 2016 forecast. These are also repeated year after year; the only source of variation is the random realization of district outcomes.

In the s-th world, we generate a sequence of 1,000 consecutive, identically distributed elections. For district-level evaluation, at any time T, we define the running metrics $\text{MAE}^{\text{di}}_{i,T,s} = \frac{1}{JT}\sum_{t=1}^{T}\sum_{j=1}^{J}\left|\hat{Y}_{i,j,t,s} - Y_{i,j,t,s}\right|$ and $\text{MSE}^{\text{di}}_{i,T,s} = \frac{1}{JT}\sum_{t=1}^{T}\sum_{j=1}^{J}\left(\hat{Y}_{i,j,t,s} - Y_{i,j,t,s}\right)^2$. Nationally, we define $\text{MAE}^{\text{ev}}_{i,T,s} = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{\text{EV}}_{i,t,s} - \text{EV}_{t,s}\right|$ and $\text{MSE}^{\text{ev}}_{i,T,s} = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\text{EV}}_{i,t,s} - \text{EV}_{t,s}\right)^2$. These running evaluation metrics are computed for each $T \in \{1, \ldots, 1000\}$. Reported probabilities are based on Monte-Carlo estimates that aggregate over $S = 40,000$ simulated worlds, or distinct sequences, unfolding in time. Specifically, at each time point, we compute the proportion of the 40,000 simulations in which oracle i currently has a winning record when compared to competitor i'. For example, after $T = 10$ elections, we estimate $\hat{\Pr}\left(\text{MAE}^{\text{di}}_{i,10} < \text{MAE}^{\text{di}}_{i',10}\right) = \frac{1}{S}\sum_{s=1}^{S}\left(\text{MAE}^{\text{di}}_{i,10,s} < \text{MAE}^{\text{di}}_{i',10,s}\right)$. This process is repeated for each competing forecaster, each evaluation metric, and each point in time.