POLITICS

Special Issue on Forecasting the 2024 US Elections

Long-Range State-Level 2024 Presidential Election Forecast: How Can You Forecast an Election When You Don't Know Who the Candidates Are Yet?

Jay A. DeSart

Utah Valley University

This model generates projections of the national popular vote and Electoral College votes a year in advance of the U.S. Presidential Election, before each party's nominees are known. It forecasts the Democratic two-party popular vote in each state and the District of Columbia. It uses four independent variables: national head-to-head polling data 13 months prior to the election, the states' prior election result, a party-adjusted home state advantage dummy variable, and a party adjusted variable simply counting the number of consecutive terms the current incumbent party has occupied the White House. New to this year's model is a polling average approach that encompasses all possible candidate matchups for whom data is available. This year's forecast suggests a distinct possibility of an Electoral College misfire benefitting the Republicans.

This is a "preproof" accepted article for *PS: Political Science & Politics*. This version may be subject to change during the production process.

DOI: 10.1017/S1049096524000805

The developments of July 21, 2024 brought one of the fundamental challenges of election forecasting into sharp focus. Forecast models that generate predictions based on the individual head-to-head matchups between candidates were dealt a significant blow when Joe Biden announced that he was withdrawing from the 2024 presidential election. This was particularly the case for my own model, which generates a forecast a year prior to election day. I had used it to generate long-range forecasts for both the 2016 and 2020 elections and had presented its preliminary forecast for 2024 at last year's American Political Science Association meeting in Los Angeles (DeSart 2023). In what now seems in hindsight to be a prescient statement, the focus of that presentation was on the challenges in generating an election forecast well before the nominees are known.

The unique contribution of this model to the forecasting literature is that it pushes the lead-time envelope by producing both national popular vote and Electoral College forecasts a year in advance of the election, long before the nominees of each party are known. This has typically necessitated generating a matrix of conditional forecasts representing the various potential matchups between each Republican candidate against each Democratic candidate.

It's a laborious process, because even though it can give us a glimpse of how each potential pairing might end up, it typically means data needs to be gathered for each matchup to generate several different point estimates. For example, in October 2015 there were six Republican candidates (Jeb Bush, Ben Carson, Ted Cruz, Carli Fiorina, Marco Rubio, and Donald Trump) and two Democratic candidates (Hillary Clinton and Bernie Sanders) for whom there was available polling data necessary for the model to generate a prediction. The first presentation of this model at the 2015 Iowa Conference on Presidential Politics (DeSart 2015) produced 11 separate forecasts (no polling data were available for the matchup between Jeb Bush and Bernie Sanders). The preliminary forecast I presented at the 2019 Annual Meeting of the American Political Science Association (DeSart 2019) featured a one by five matrix showing conditional forecasts pitting the incumbent Donald Trump against five potential challengers (Joe Biden, Bernie Sanders, Elizabeth Warren, Kamala Harris, and Pete Buttigieg).

Biden's surprise withdrawal late in the campaign season this year highlights the necessity of taking this broad approach if one wants to cover the multitude of possibilities that might develop over the next several months. However, this problem isn't unique to the unusual circumstances of the current campaign. Michael Bloomberg's surge in the polls in December 2019, a month after I announced my long-range forecast for 2020, threatened to render the forecast meaningless because pollsters hadn't even begun treating Bloomberg as a potential nominee until December.

It was this particular challenge, not knowing for sure who the nominees will be that far in advance, that prompted a different approach that I proposed a year ago: averaging the available polling data across the potential matchups in order to generate a single point estimate.

The Long Range State-Level Forecast Model

This model (DeSart 2015; 2019; 2021) generates state-level popular vote forecasts in each of the 50 states and the District of Columbia. These predictions can then be extrapolated to national level forecasts by awarding each state's electoral votes to the predicted popular vote winner, and by calculating a turnout-weighted average of each state's popular vote forecast to generate a national popular vote projection.

The state-level forecasts are the prediction of the Democratic share of the two-party popular vote in each state as a function of four variables:

PRIOR RESULT - the share of the two-party popular vote won by the Democratic candidate in each state in the previous presidential election.

POLLS - the average Democratic two-party share of national head-to-head polls taken 13 months in advance of the election, in October of the year prior to the election.

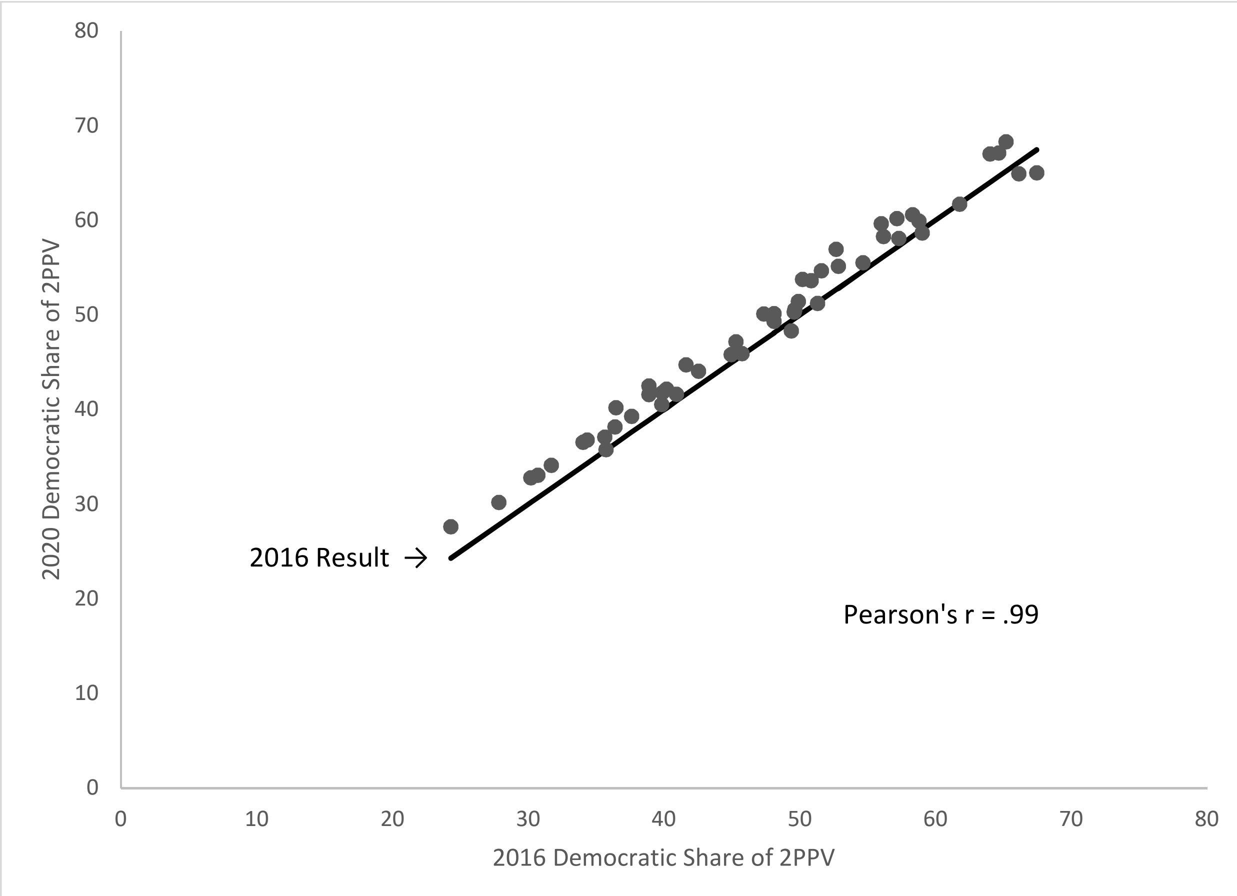
HOME STATE - a dummy variable for each state indicating its status as a home state for each candidate, signed according to party (i.e., positive for a Democrat and negative for a Republican). Due to the inclusion of the lagged dependent variable as a predictor variable, it is also necessary to have an opposite signed value for the party-adjusted home state dummy variable from the previous election to account for the removal of previous candidates' advantages.

CONSECUTIVE TERMS - a simple party-adjusted variable (i.e., positive for Democrats and negative for Republicans) that captures the number of terms a party has consecutively occupied the White House going into the election.

Not surprisingly, the lagged dependent variable as a predictor dominates the model as their bivariate correlation is quite strong (Pearson's r = .90). States' relative positions with each other in terms of the partisan distribution of their election results do not typically shift much from one election to the next as demonstrated in Figure 1.

Despite the dramatic difference in the outcomes from 2016 to 2020, the Democratic share of the two-party popular vote (D2PPV) was remarkably stable over that time, with a Pearson's r of .99. The line in Figure 1 represents the pattern if the 2016 results had been a perfect predictor of those of 2020. While it is clear that the correlation is very strong, there is still a fairly systemic shift from 2016 to 2020. It's that systemic shift that the model is ultimately trying to capture.

FIGURE 1: The Correlation Between 2016 and 2020 State-Level Presidential Election Results



Of course, the challenge is the availability of suitable predictor variables that far in advance of the election that can explain that shift. The main predictor designed to do that is the TERMS variable. The two-term penalty is a now well-documented phenomenon in presidential elections. Norpoth (1995) pointed out the cyclical pattern in the popular vote outcomes of presidential elections across time. Abramowitz (1988) employs a two-term penalty term in his Time-for-Change model. The consecutive term count variable in this model attempts to capture this cyclical shift from one election to the next.

The Challenge of Pre-Nomination General Election Forecasts

The biggest challenge in attempting to generate a forecast this far in advance of the election is that we don't yet know who the nominees will be. Given that two key variables in the model, POLLS and HOME STATE, are dependent upon the specific matchup between candidates, not knowing which candidates will face-off against each other a year later complicates matters. The process of generating this forecast involves calculating the polling averages for each matchup, generating the state-level popular vote predictions (while accounting for the home-state advantages of each candidate), extrapolating the state-level predictions to national-level outcome by calculating a turnout-weighted average for a national popular vote projection, and awarding Electoral College votes to the candidates based on the state popular vote forecast, and then finally running Monte Carlo simulations to calculate state and national win probabilities.

The larger the field, the more time consuming this process becomes. There are two potential solutions to this problem. One is to simply make a judgement on which candidates are the two most likely to win their parties' nominations. That might be easier in some years rather than others. When an incumbent president is running for reelection that typically narrows down the field on one side of the ballot but can still produce a long list of potential challengers. Even so, the developments of 2024 show that even when an incumbent is running for reelection there's no guarantee that they will eventually be the nominee.

Choosing among a field of potential candidates who are considered the "most likely" to win the nomination is a challenge in and of itself that far in advance of the election. In November 1991, the clear frontrunner in polling for the Democratic nomination was California Governor Jerry Brown. The eventual nominee, Arkansas Governor Bill Clinton, was still trailing behind

Brown, Iowa Senator Tom Harkin, and Nebraska Senator Bob Kerrey at that point. Generally speaking, the media outlets and polling organizations have done a fairly good job of assessing the field of candidates early on when deciding which candidates are "top-tier" and which ones will likely be considered "also rans." Up to this point it has been relatively easy to find the necessary head-to-head polling data between the two eventual nominees 13 months in advance of the election in order to produce a forecast. Even so, it is easy to think of a scenario where the list of declared candidates thirteen months ahead of the election might not include the eventual nominee.

Given that the main goal of this model is to *accurately* project an outcome of the election a year in advance and not leave out potential candidates, one either needs to generate a full matrix of all possible matchups for which there is available data, or to come up with a way of producing a single point estimate that shows the likelihood of one party's nominee winning over the other. In 2016 and 2020, I opted for the former approach, with all of the work and challenges that it entails. Fortunately, those matrices have included the eventual matchup.

While there is some variation in the projections generated across the different matchups, they generally do tend to point in the same direction with a few exceptions. Matchups with lesser-known candidates (who often end up dropping out) tend to have much less available polling data and tend to have closer margins and higher proportions of undecided respondents. While the variation across projections gives us a glimpse of the potential impact that candidate characteristics (at least those that are known that far in advance of the election) may have on the outcome, it may be of greater interest to focus on the broader partisan context that the candidates will face on Election Day, and whether or not we can capture that context so far in advance of the election.

A Data Averaging Approach

If the intent is to capture the overall context rather than the specific matchups, then it may be beneficial (and less work) to average the available data instead of trying to generate a forecast for each possible matchup. Apart from the amount of work involved in producing multiple forecasts across the multiple candidate pairings, there is also an issue regarding the reliability of those forecasts that rely upon a relatively small number of polling data points. Fortunately, there were plenty of polls conducted 13 months before the general election that asked respondents to choose between Hillary Clinton and Donald Trump in 2015, and between Joe Biden and Donald Trump in 2019. However, for example, if the 2016 matchup had turned into a race between Bernie Sanders and Ted Cruz, or between Hillary Clinton and Mike Huckabee, the forecasts for those contests would be based on a single survey, and therefore would be highly dependent upon any sampling issues present in that one poll. As it was, there were 10 polls conducted in October 2015 that pitted Hillary Clinton against Donald Trump. Therefore, the sampling errors in any one of those polls would potentially be ameliorated by averaging it with the others.

Poll aggregation is a widespread practice among those who report poll results. RealClearPolitics, FiveThirtyEight, and the now defunct Pollster.com all employ an averaging technique to estimate the "true" population parameter of a variety of survey questions, not just election polls. Several forecast models employ the strategy of averaging polling data as one of their predictor variables (DeSart and Holbrook 2003; Graefe 2018; Graefe et al. 2014; Holbrook 2008; 2012). The central-limit theorem in probability theory suggests that the mean of a sampling distribution should be equal to the true population mean, assuming said distribution is made up of sample statistics from unbiased probability samples. (Billingsly 1995)

The approach I am using for the 2024 forecast extends that principle to averaging the polls not just within each matchup, but across *all possible matchups*. The result of such an approach should yield a measure of the overall partisan context going into the election by not just balancing out the random sampling errors, as suggested by the central-limit theorem, but will also mute the candidate-specific effects that each particular matchup brings to each specific poll result. However, given that polling organizations generally tend to ask more frequently about top-tier candidates much more often than the also-rans, greater weight will be given to the poll results featuring candidates that are most likely the actual candidates who will face each other in the general election. The presence of the polls featuring lesser-known candidates will likely moderate the overall mean to capture something closer to the general partisan context of the election at large.

The inclusion of polling data as an independent variable does introduce a potential source of prediction error into the model. Given the apparent polling misfires in 2016, one might question the efficacy of adding polls as a predictor. Its contribution to the prediction can be mitigated by polling error, especially if there is systemic error in the polls that under- or over-estimate the support for a given candidate.

That concern should be alleviated by two things. First, while polling error is clearly a concern, the direction and extent of that error varies from one election to the next, and the average polling error has generally declined over time. While the average error in 2020 was slightly higher than that in 2016, it was still on par with polling errors since the 1960s (Clinton et al. 2021). Furthermore, the polling misfires in 2016 were mostly present at the state-level, while the national-level polls performed quite well (Kennedy et al. 2017), and this model relies upon national polls. Second, the POLLS variable performs quite well in the model achieving statistical significance in the hypothesized direction. Ultimately, despite any potential issue for polling error in any given election, the polls do provide a useful contribution to the explanation of election results over time.

Allocating Home State Advantage

One problem created by averaging the polling data across matchups is that we lose the specific nature of determining the home-state advantage in each candidate pairing. When you have a head-to-head forecast it's simply a matter of assigning the party-adjusted Home State dummy variable to each candidate's home state. If all polls across all matchups are averaged together, we lose the ability to assign that dummy variable to any specific state.

To get around that issue, I am choosing to allocate the Home State advantage proportionally across the field of candidates from each party. Instead of assigning a value of 0 or 1, I allocate a value to the Home State variable equal to the proportion of time each candidate appears in all of the polls from that candidate's party. For example, in all 123 of the polling matchups conducted in October 2019, Bernie Sanders was listed as the Democratic candidate in 20 of them. Therefore, under the proposed approach, a value of .162 (20/123) would be given to the Home State variable for Sanders' home state of Vermont. On the other hand, Delaware would be given a value of .407 to account for the fact that Joe Biden was the Democratic candidate in 50 of those matchups[1].

Table 1 below shows the impact that using this approach would have had on the model's *a priori* forecasts in 2016 and 2020. In each instance, the national popular vote projection in each forecast was improved over that using only the polls featuring the two eventual nominees in each election. This suggests that an averaging approach would tend to improve the overall performance of the model moving forward.

Table 1: Performance of Polling Averages

Year	Result (National D2PPV%)	Matchup			Averaged Across Matchups		
		Poll Average	Forecast	Forecast Error	Poll Average	Forecast	Forecast Error
2016	51.1	50.7	50.3	-0.8	51.0	50.5	-0.6
2020	52.3	54.9	54.8	+2.3	52.8	51.6	-0.7

The Model for 2024

With those modifications, along with incorporating the data from 2020 to update the model, the coefficients I used to generate the forecast for 2024 can be found in Table 2 below. All four variables remain statistically significant with the inclusion of the observations from 2020. In addition, the values of their coefficients remained fairly stable compared to those used in previous elections.

Using these coefficients and the new approach I am proposing here, I generated a forecast of the 2024 presidential election in November 2023. Using the RealClearPolitics Election Polls archive I obtained a total of 85 general election polling matchups from October 2023. Joe Biden was listed as the Democratic candidate in 80 of these polls and Donald Trump was listed as the Republican candidate in 68. Joe Biden and Donald Trump were featured as the matchup in 65 of the 85 polls. That The remainder featured various matchups between either of these candidates with potential opponents, including Bernie Sanders, Kamala Harris, Nikki Hayley, Ron DeSantis, and Mitt Romney. The distribution of the matchups, and the resultant impact on calculating the potential home state advantage variable is presented in Table 3.

Table 2: Updated Long-Range State-Level Forecast Model

Independent Variable	Unstandardized Regression Coefficient	Standard Error
Prior Result	1.017	.018
Previous October Polls	0.524	.051
Home State Advantage	2.544	.747
Number of Terms	-0.988	.136
Constant	-27.423	2.785

R^2 = .90

S.E. $_{y|x}$ = 3.08

N = 350

In-Sample Model Performance over Time

	1996	2000	2004	2008	2012	2016	2020	OVERALL
States Correctly Predicted	88%	84%	94%	92%	100%	90%	96%	92%
Mean Absolute Error	2.74	2.57	2.01	2.75	1.88	2.56	1.47	2.28
National-Level Predictions (Excluding DC)								
National Popular Vote	53.6	48.6	49.7	52.7	52.5	49.8	54.1	
Error	-1.1	-1.6	+1.0	-0.9	+0.6	-1.2	+1.8	1.2†
Electoral College	399	231	228	333	329	269	347	
Error	+23	-33	-21	-28	0	+39	+44	27†

† Mean Absolute Error

Table 3: Distribution of Candidate Appearances in Polls and Resultant Home-State Advantage Variable for 2024

Party	Candidate	Home State	Number of Polls	Home State Advantage Dummy
Republican	Donald Trump	Florida	65	-.953
	Ron DeSantis	Florida	12	
	Nikki Hayley	South Carolina	4	-.047
	Mitt Romney	Utah	1	-.011
Democrat	Joe Biden	Delaware	80	-.059
	Kamala Harris	California	4	.047
	Bernie Sanders	Vermont	1	.011

Given the way the home state advantage variable is coded, along with the lagged result variable as a predictor, the advantage is already present in the data for the incumbent president running for reelection. This would result in a value of 0 for Delaware in 2024. However, since a handful of polls did not feature Biden as a candidate, a slight adjustment is made to account for that. In 2020, Donald Trump was still considered as a New York resident, so the Florida advantage still needed to be accounted for in the 2024 forecast. In addition, the impact of the presumed New York advantage for Donald Trump in 2020 is accounted for by coding New York with a value of 1 in the home state advantage variable. This would represent the return to "normal" for New York in 2024.

Given the new coefficients in Table 2, these values for the Home State variable in Table 3, and using polling data from October 2023, the model generates a forecast of the 2024 election which suggests that an Electoral College misfire is a distinct possibility. The model projects that the Democratic candidate will win the national two-party popular vote 50.7% to 49.3%.

However, when I use the forecast's state-level point estimates and simply award each state's Electoral Votes to the candidate forecasted to win a majority its two-party popular vote, the projected Electoral College has the Republican candidate winning a majority 226 to 312.

Table 4 shows the results of the Monte Carlo simulations in which 100,000 elections were generated with the model's state-level predictions, while allowing them to randomly vary in a normal distribution around that point estimate using the model's standard error of the estimate, 3.08. The mean of the distribution of these simulated election results represents the model's forecasts of both the national popular vote and Electoral College outcomes.

Table 4: Monte Carlo Simulation Results (2024 Forecast)

	Democratic Candidate	Republican Candidate
Mean Projected share of National Two-Party Popular Vote	50.7%	49.3%
95% Confidence Interval	49.4% - 51.9%	48.1% - 50.6%
Mean Projected Electoral College Vote Total	256	282
95% Confidence Interval	218 – 306	232 - 320

Table 5 presents the distribution of outcomes in these simulated elections, based on which candidate wins a majority of the national two-party popular vote and which candidate wins an Electoral College majority. You can see from these results that the model suggests that another Electoral College misfire is a distinct possibility. The Democratic candidate won the national popular vote in 86% of the 100,000 simulated outcomes. However, the Democratic candidate won an Electoral College majority in roughly a quarter (25.2%) of the simulated elections. This means that these projections suggest a 61% chance of a repeat of the elections of 2000 and 2016 wherein the Republican lost the popular vote but won a majority in the Electoral College. Overall, these results suggest that the Republicans have around a 74% chance of regaining the White House as a result of this year's election.

Table 5: National Popular Vote and Electoral College Vote Outcomes in 100,000 Simulated 2024 Presidential Elections

		National Popular Vote Result	
		Republican Wins	**Democrat Wins**
Electoral College Result	**Republican Wins**	**12.8%**	**60.9%**
	Tie	**0.1%**	**1.0%**
	Democrat Wins	**1.0%**	**24.2%**

Concluding Thoughts

At the very least, it should be clear that the model suggests that the election will be very close, perhaps closer than it was in 2020. However, we must recognize that we are dealing with an unprecedented set of circumstances this year. Biden's late departure only solidifies the need to move beyond a model that is specifically tied to individual matchups. Trump's felony convictions earlier this year also injected a degree of potential uncertainty as well. This degree of uncertainty is what makes election forecasting challenging, especially a year in advance of the election. Even so, this model has done reasonably well in capturing the systematic shifts from one election to the next, even when the specific candidates who will eventually appear on the ballot is in doubt. It's entirely possible than in any given election that the candidates appearing on the ballot in November were not even under consideration in the polls 13 months prior to the election. So an averaging approach that far in advance should help mitigate some of those elements of uncertainty. While it may work to mute the candidate-specific factors that may affect the outcome, it allows us to gauge the overall partisan context underlying the dynamics of the campaign… especially when the specific candidates themselves are somewhat in doubt.

Admittedly, the rather wide confidence interval on the Electoral College projection presented in Table 4 doesn't instill a great deal of comfort in the reliability of the model. The conclusion we could draw from that is that the model suggests that either candidate has a not insignificant chance of winning. That being the case, it is reasonable to ask whether this model is worthy of attention at all.

Therein lies the downside of a long-range forecast. In that regard, it is not unlike hurricane forecast models that have an ever-widening cone of uncertainty the longer the time frame of the prediction. Table 1 demonstrates that the projections tend to center around the actual result with a reasonable amount of variability and that they do generally tend to point in the "correct" direction. Even so, the fact that the model's projections for this year's election leaves so much in doubt is testimony to the level of uncertainty that voters may have had in October 2023. Biden's withdrawal in July appears to have removed a large source of that uncertainty and more recent polling data suggests a much more favorable context for the Democrats than this model's projections suggest. Ultimately, the results of the election will issue a verdict on this model’s efficacy and the utility of employing such a long lead-time in predicting the outcome of elections.

Data Availability Statement

Research documentation and data that support the findings of this study are openly available at the Harvard Dataverse at https://doi.org/10.7910/DVN/HZGLY9

Conflicts of Interest

The author declares no ethical issues or conflicts of interest in this research.

References

Abramowitz, Alan. 1988. An Improved Model for Predicting Presidential Election Outcomes. *PS: Political Science and Politics*, 21: 843–847.

Billingsley, Patrick. 1995. *Probability and Measure.* 3rd Edition. Hoboken, NJ: John Wiley and Sons.

Clinton, Josh, Jennifer Agiesta, Megan Brenan, Camille Burge, Marjorie Connelly, Ariel Edwards-Levy, Bernard Fraga, Emily Guskin, D. Sunshine Hillygus, Chris Jackson, Jeff Jones, Scott Keeter, Kabir Khanna, John Lapinski, Lydia Saad, Daron Shaw, Andrew Smith, David Wilson, and Christopher Wlezien. 2021. *Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Poll.* AAPOR.

DeSart, Jay. 2015. “State Electoral Histories, Regime Age, and Long-Range Presidential Election Forecasts: Predicting the 2016 Presidential Election.” Paper presented at the Iowa Conference on Presidential Politics. Sauk Center, IA.

DeSart, Jay. 2019. “A Long-Range State-Level Presidential Election Forecast Model in 2016 and 2020.” Paper presented at the Annual Meeting of the American Political Science Association. Washington, DC.

DeSart, Jay. 2021. “A Long-Range State-Level Forecast of the 2020 Presidential Election.” *PS: Political Science and Politics*. 54: 73-76.

DeSart, Jay. 2023. "’But You Don't Even Know Who the Nominees Are Yet!’ - Preliminary 2024 Forecast.” Paper presented at the Annual Meeting of the American Political Science Association, Los Angeles, CA.

DeSart, Jay, 2024, "Replication Data for: Long-Range State-Level 2024 Presidential Election Forecast: How Can You Forecast an Election When You Don't Know Who the Candidates Are Yet?" *PS: Political Science and Politics.* https://doi.org/10.7910/DVN/HZGLY9

DeSart, Jay and Thomas Holbrook. 2003. "Statewide Trial Heat Polls and the 2000 Presidential Election: A Forecast Model." Social Science Quarterly. 84: 561-573.

Graefe, Andreas. 2018. "Predicting elections: Experts, polls, and fundamentals." *Judgement and Decision Making.*

Graefe, Andreas, J. Scott Armstrong, Randall J. Jones, and Alfred G. Cuzán. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting.* 30: 43–54.

Holbrook, Thomas M. 2008. "Incumbency, National Conditions, and the 2008 Presidential Election." *PS: Political Science and Politics.* 41: 709-712

Holbrook, Thomas M. 2012. "Incumbency, National Conditions, and the 2012 Presidential Election." *PS: Political Science and Politics.* 45: 640-643.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Doug Rivers, Lydia Saad, Evans Witt, and Chris Wlezien. 2017. *An Evaluation of 2016 Election Polls in the United States.* AAPOR.

Norpoth, Helmut. 1995. "Is Clinton Doomed? An Early Forecast for 1996." *PS: Political Science & Politics* 28: 201–7.

[1] One might question whether this approach of averaging the Home State Advantage is completely necessary, as it could often result in a relatively negligible value for the variable for those candidates who appear in very few polling matchups. Even so, in order to hold true to the spirit of the data averaging approach, I deemed it necessary to keep the model internally consistent.