# Dynamic Hierarchical Modeling for U.S. Presidential Forecast

Poetto Patrick - 5201214, Manuel Bottino - 5201213

2024-10-25

1 Introduction

The United States presidential rush represents the most important political event worldwide. Every four years, the two major parties primary winners compete for 538 voters in a single round with a clear majority. The candidates who gets to win at least 270 votes gains the Presidential mandate.

(a) General Elections Map

(b) 2020 Final Results
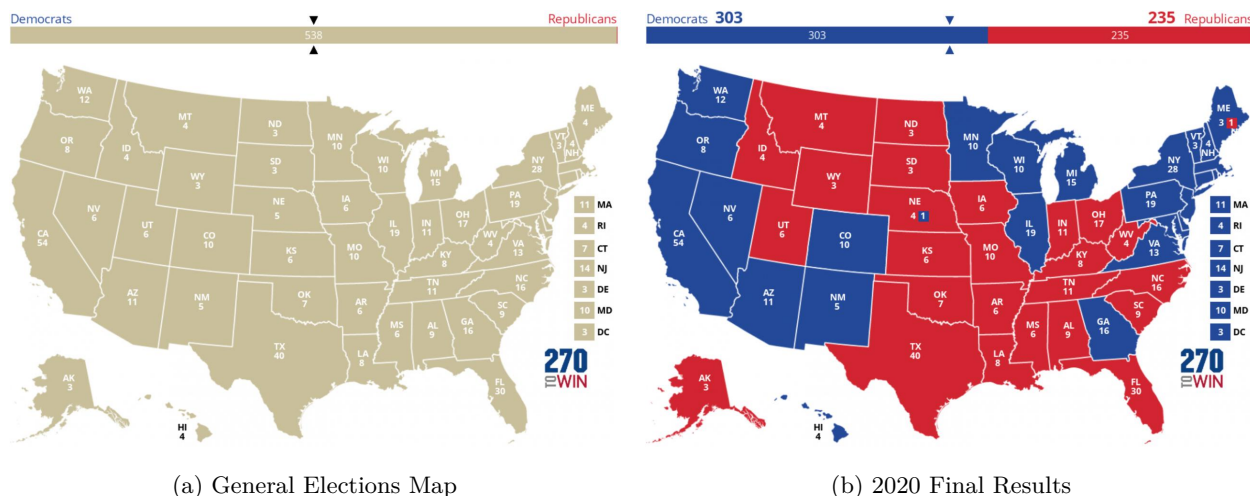
Despite the wider and wider data availability and accessibility, forecasting the outcome of such complex events as modern election cycles is still a tough task as eminent precedents such as the disastrous failure of the 2016 presidential forecasts (Why 2016 election polls missed their mark) and the against-all-odds outcome of the UK Brexit referendum (Here's why the majority of Brexit polls were wrong) have shown in the close past.

This project is aimed at modeling vote intentions at the day of the elections in order to forecast the 2024 new President. The present work bases on (Linzer 2013) proposal which has been then updated by (Heidemanns, Gelman, and Morris 2020) and tries to implement the modeling procedure closely following the one used by "The Economist" journal (How The Economist's presidential forecast works) for the 2020 and 2024 rushes, which implementation is available at this link. The vote intention forecasts for Election Day are conducted following a Bayesian dynamic hierarchical structured model by combining fundamentals and trial-heat polls (see section 2) as predictors. As stated by (Heidemanns, Gelman, and Morris 2020), while fundamentals such as economic growth and presidential approval are highly predictive early in the electoral cycle due to electors tendence in evaluating whether the party in charge is doing well or not, Trial-heat polls become more predictive as we approach Election Day since there is a natural convergence of intentions toward actual vote as Election day gets closer (e.g. if one elector has a specific vote intention, that intention will be much less likely to change two days before Election day rather than six months before).

2 Data and methods

2.1 Data

All the data used to fit the model comes from two main sources:

- Polls: trial-heat results coming from different pollsters both nationwide and statewide at different times: different bias terms for this kind of records will arise since despite the usual statistical procedure to ensure the consistency of the randomness assumption, achieving such a goal is almost impossible due to different hard to abate biases. Modeling of biases and variance will be faced more specifically in the following section, taking into account the fact that different pollsters tend to favor one candidate over the other, the survey mode has a specific impact on the results collected, and most importantly, correction terms for differential nonresponse rates have to be included. This last issue became a big deal since in the close past most forecasts have failed supposedly due to it. People who are contacted for the poll and decide not to participate represent a latent effect that must be taken into account, specifically if they can be identified as Republicans or Democrats. The difference between the former and latter in the sampled population who doesn't respond is meaningful in terms of forecast correction.

- Social and economic data: this category does include geographical features that are proposed by means of variance-covariance matrices that represent the spatial correlation of the population features, which is clearly disclosed by looking at any election map where most of the times one may notice the west and east coasts are usually blue (democrats) while the inland has in general a republican prevalence. In principle, a spatial correlation must be somehow taken into account according to Tobler's first law of geography. Other meaningful data are recorded to keep track of social features of the population like the percentage of white evangelics rather than other ethnicities prevalence in specific regions, the education rates and the per capita GDP.

Data source: https://projects.fivethirtyeight.com/polls/president-general/2024/national/

2.2 Data processing

In this section the data pre-processing procedure implemented to wrangle the data is presented (see the code for the details). First of all, the Stan environment is set up with specific parameters, including the number of cores, chains, and sampling periods, to optimize the Bayesian computation (see the implementation paragraph). The primary task is to preprocess and clean polling data from two sources: "all_polls_2012.csv" and "president_polls_2024.csv". This involves reading in the datasets, standardizing state names, and selecting relevant columns. The data is then reshaped to create one column per candidate. Additional columns are computed to represent the sum of non-important candidate percentages (others) and the percentage of undecided voters. Missing values are handled with zeros to ensure data completeness and consistency. Subsequently, the code defines three essential functions for further analysis. The cov_matrix function generates a variance-covariance matrix, which is crucial for understanding the relationships between variables. The mean_low_high function computes the mean and confidence intervals for the draws, providing a statistical summary of the data. The check_cov_matrix function is used to validate the covariance matrix by printing histograms and calculating the national standard deviation. In the data wrangling section, the code selects relevant columns from the processed polling data and performs basic data transformations. It handles missing values and inconsistencies in pollster names, and computes vote shares for the candidates. Numerical indices for states, days, weeks, and pollsters are created to facilitate further analysis. The code then retrieves state contextual information from 2012 data, including state names, electoral votes, and prior differences in scores. It calculates state weights based on the share of the national vote and creates covariance matrices. Various demographic and socioeconomic variables are merged from different sources, including census data, urbanicity indices, and the percentage of white evangelical populations. The data is normalized and transformed into a wide format for easier manipulation and analysis.

2.3 Visualization

As stated in the previous paragraph, the hierarchical model mainly combines trial heat polls and socio economic data. Some visualization to get an idea of how do they look like is proposed in the following with no aim of completeness. Data visualization in this setting may not be so meaningful since most of the data has the very same structure (as all the polls) or may be already well known (economical features of the United States). Only a couple of example to get an idea of the main structure are proposed.
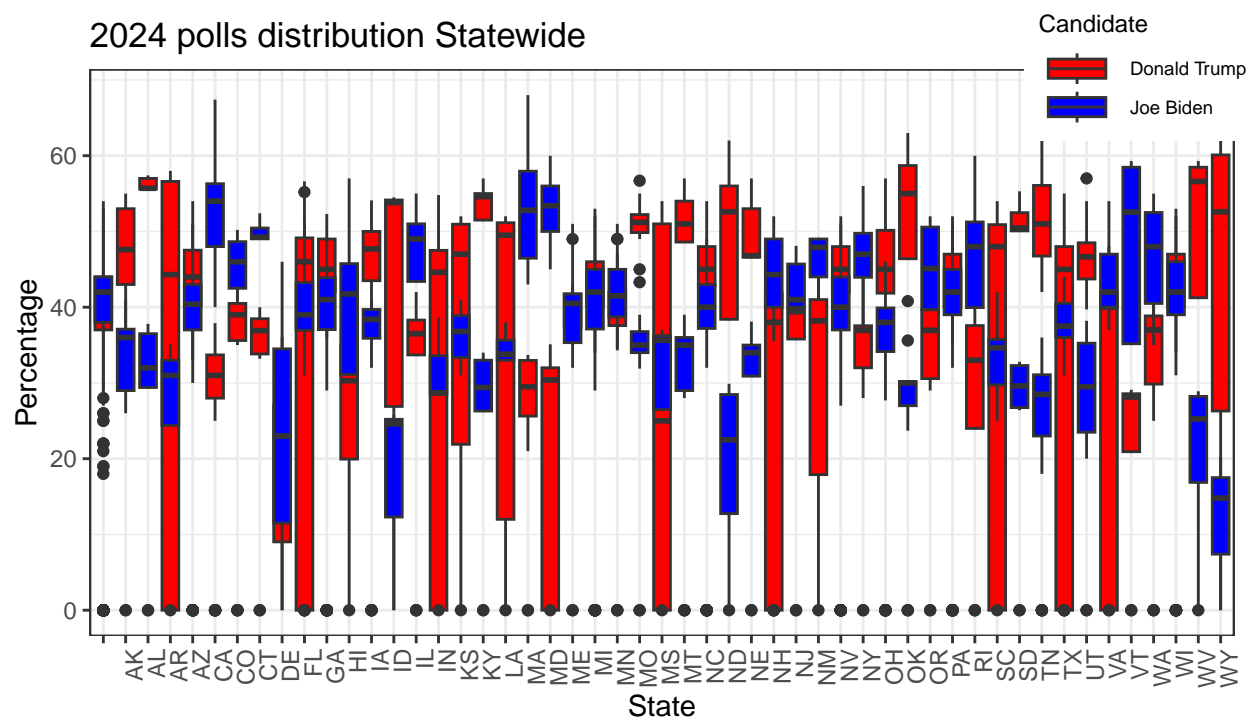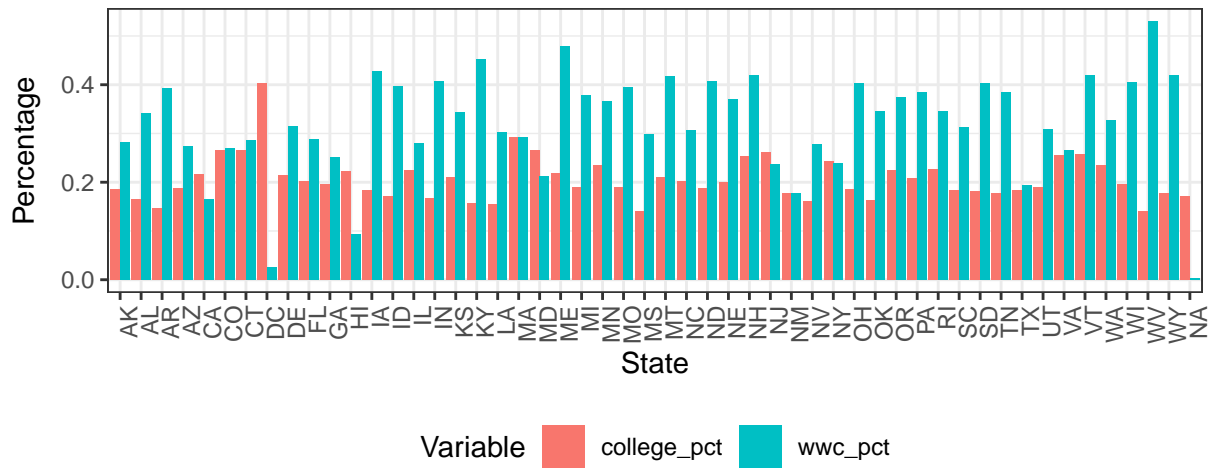


```
## 'geom_smooth()' using formula = 'y ~ x'
```
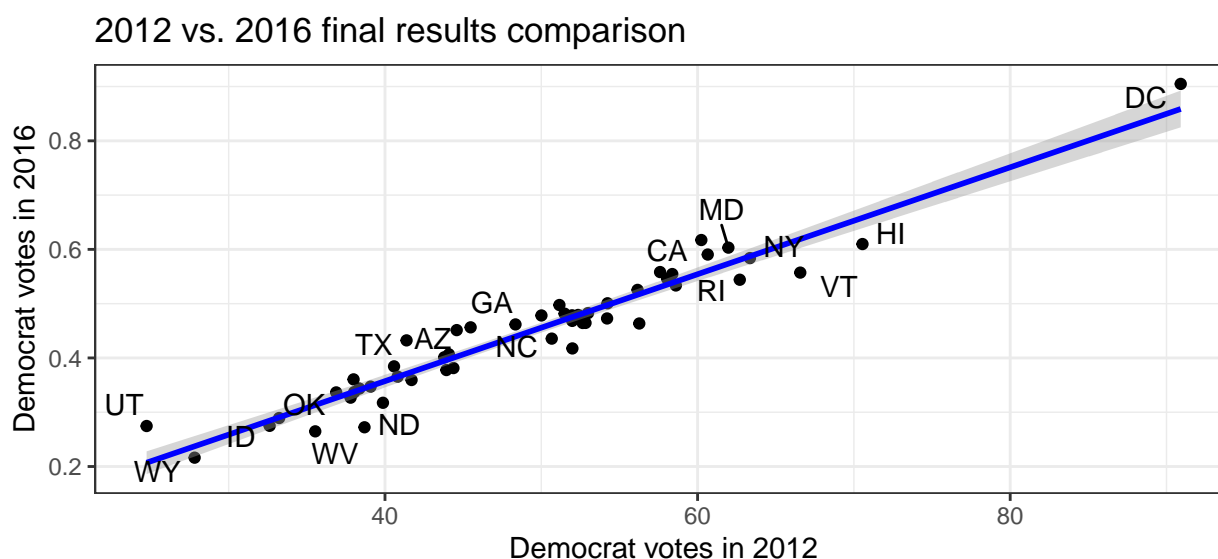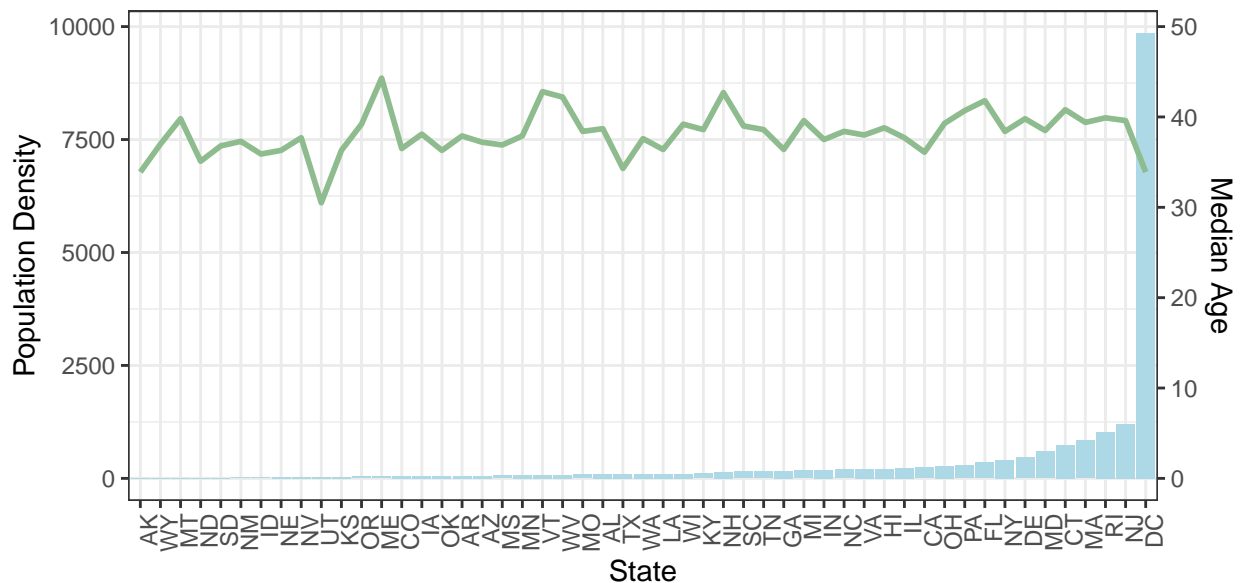
Ethnic prevalence statewide

Ethnicity: black_pct, hisp_other_pct, white_pct
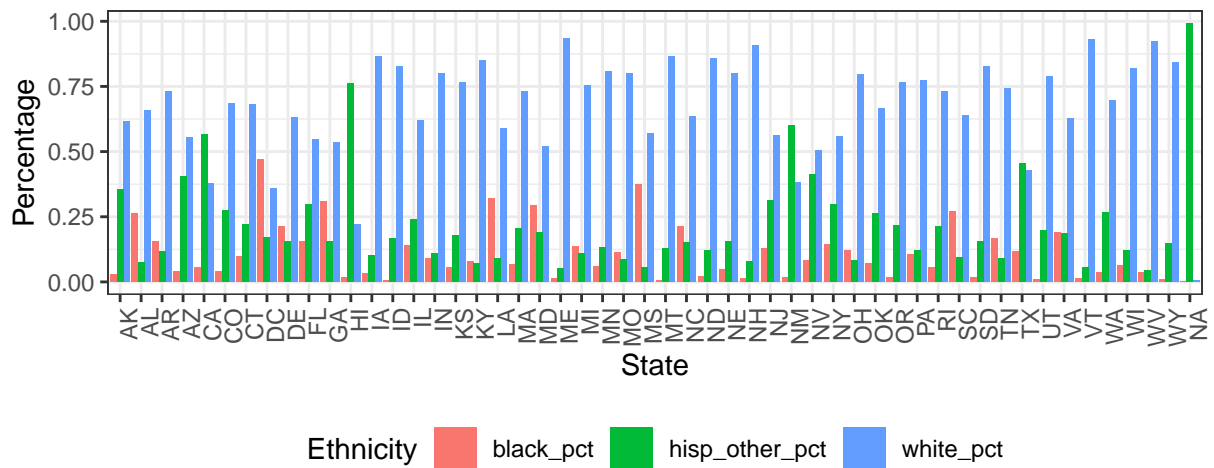
College graduated and white working class statewide

Variable: college_pct, wwc_pct

Society features

2.4   Methods for inference approximation - STAN Package

The Stan package will be used to approximate Bayesian inference in the following sections. This package allows for the analysis of complex statistical models using Bayesian approaches, which combine prior information with observed data to update the probability distributions of model parameters. Stan employs the Hamiltonian Monte Carlo (HMC) method to efficiently sample from posterior distributions, overcoming the limitations of traditional methods like Metropolis-Hastings. This enables accurate and reliable estimates even for high-dimensional models. The specifications that have been set at beginning of this section allow to enhance parallel computation.

The Hamiltonian Monte Carlo is a method that improves sampling efficiency by using the geometry of the parameter space:

1. Hamiltonian Dynamics: HMC treats the target distribution as a physical system where each parameter θ represents a position in space, and an auxiliary variable p (momentum) is introduced. The joint density of (θ, p) defines a Hamiltonian $H(\theta, p) = U(\theta) + K(p)$, where:

   - $U(\theta)$ is the potential energy, given by the negative log probability of the target distribution.
   - $K(p)$ is the kinetic energy, typically chosen as $K(p) = \frac{1}{2}p^T M^{-1} p$, where M is a mass matrix.

2. Leapfrog Integrator: To simulate the Hamiltonian dynamics, HMC uses the leapfrog integrator, which approximates the continuous trajectories by alternating small steps for updating positions and momenta. This ensures volume preservation and reversibility.

3. Proposal Generation: Starting from the current state (θ, p), the leapfrog integrator generates a proposal by simulating the dynamics for a specified number of steps and step size. The proposal (θ', p') is then evaluated.

4. Metropolis-Hastings Correction: To ensure the correct target distribution, the Metropolis-Hastings acceptance step is applied to the proposal. The acceptance probability is determined by the change in Hamiltonian, ensuring detailed balance.

To set up a model in Stan, a block structure is required, where in each block a specific feature is defined (e.g. data, parameters ecc.) and eventually a model block is defined along with priors and likelihood to run the model. Other blocks may then be defined to retrieve quantities derived after the sampling.

2.5   Model

Following (Linzer 2013), (Heidemanns, Gelman, and Morris 2020) set up a model that generates a comprehensive Election Day forecast by integrating two primary sources of information: historical analysis incorporating economically and politically significant variables (such as personal income growth, presidential approval ratings, and incumbency status) and real-time data from state and national polls conducted during the election season. These two streams of data are synthesized through a time-series model that evaluates state and national opinion trends. Additionally, the model addresses certain aspects of non-sampling errors inherent in polling methodologies. This combination of the two sources of information in a model, also known as fundamentals-based model, was first introduced by (Abramowitz 2008) with the "Time-for-Change" model, a forecasting approach developed by political scientist to predict the outcomes of U.S. presidential elections. This model is based on three key indicators that are known long before Election Day:

1. Economic Growth: The growth rate of the economy during the second quarter of the election year, and the economic conditions significantly influence voting behavior, with voters likely to reward the incumbent party during periods of economic growth and penalize it during economic downturns.

2. Presidential Approval: The incumbent president's approval rating at mid-year.

3. Time-for-Change Factor: The length of time the incumbent president's party has controlled the White House. There is an electoral fatigue factor where voters tend to prefer a change after a party has been in power for two or more terms. This reflects a desire for periodic alternation in political leadership.

In this sense one novelty introduced by (Heidemanns, Gelman, and Morris 2020) is an interaction term between the economic growth and the share of swing voters in the electorate, as measured by American National Election Study.

2.5.1 Polls Data

Each poll provides an estimate of the day's average support for the two main party candidates. With y_i representing the number of respondents supporting the Democratic candidate in poll i, and with n_i being the total number of respondents supporting one of the major party candidates. We start with the binomial sampling model:

$$y_i \sim \text{Binomial}(\theta_i, n_i) \tag{1}$$

- States pools are modeled as:

$$\theta_i = \text{logit}^{-1}\left(\mu^b_{s[i]t[i]} + \alpha_i + \zeta^{\text{state}}_i + \xi_{s[i]}\right) \tag{2}$$

It includes terms for state-level support $\mu^b_{s[i]t[i]}$, which represents the underlying support for the s, t Democrat in state s at time t, shared bias α_i, measurement error ζ^{state}_i, and state-level error $\xi_{s[i]}$. These biases arise because opinion polls are not random samples of the voting population.

- National pools are modeled as:

$$\theta_i = \text{logit}^{-1}\left(\sum_{s=1}^{S} w_s \mu^b_{st[i]} + \alpha_i + \zeta^{\text{national}}_i + \sum_{s=1}^{S} w_s \xi_s\right) \tag{3}$$

National polls are modeled similarly to state ones, but with weighted averages of state-level parameters w_s, that sum to 1 and which are proportional to the number of votes in the state in the previous election. It is important to notice that this model does not account for third parties.

The shared bias α_i contains adjustments for house effects $\mu^c_{p[i]}$, polling population effects $\mu^r_{r[i]}$, and mode effects $\mu^m_{m[i]}$ to correct potential biases.

$$\alpha_i = \mu^c_{p[i]} + \mu^r_{r[i]} + \mu^m_{m[i]} + z\epsilon_{t[i]} \tag{4}$$

an adjustment trend term for nonresponse bias ϵ and an indicator z equal to 1 if the pollster does not adjust for partisanship and 0 otherwise,

```r
# create priors ---------------------------------------------------
# read in abramowitz data
prior_in <- read_csv("data/state_priors_08_12_16.csv") %>%
  filter(date <= RUN_DATE) %>%
  group_by(state) %>%
  arrange(date) %>%
  filter(date == max(date)) %>%
  select(state,pred) %>%
  ungroup() %>%
  arrange(state)
```

```
mu_b_prior <- logit(prior_in$pred + 0.00)
names(mu_b_prior) <- prior_in$state
names(mu_b_prior) == names(prior_diff_score) # correct order?

national_mu_prior <- weighted.mean(inv.logit(mu_b_prior), state_weights)
```

<div align="center">

Prior Biden two-party vote is 0.509
With a national standard deviation of 0.033

</div>

2.5.1.1 State-level trends At the state-level the author models the development of public opinion as a correlated random walk, using prior information from fundamentals-based predictions on Election Day. This approach allows for the sharing of information across states and over time. The random walk component connects days with polls and interpolates for days without polls, with uncertainty being a function of the prior model's encoded public opinion change and polling data from other states.

The model assumes that similar states will exhibit similar trends in the absence of specific data, which is enforced through a correlation matrix. This matrix, derived from past election results and other state-level predictors like education, is used to ensure reasonable results for national and state-level swings. Off-diagonal elements smaller than zero are set to zero, and the matrix is scaled to achieve the desired degree of day-to-day change.

$$\mu_t^b | \mu_{t-1}^b \sim \text{MVN}(\mu_{t-1}^b, \Sigma^b) \tag{5}$$

where μ_t^b is the state-level support estimate, evolving as a multivariate normal distribution around the previous day's estimate, with covariance Σ^b. This structure allows the model to capture and adjust for state-level opinion dynamics, providing a comprehensive and adaptive framework for election forecasting.

2.5.1.2 House, population and mode effect House Effects: House effects refer to systematic biases that arise from the methodology and practices of different polling organizations. Each polling firm might have distinct approaches in terms of sampling, question phrasing, and weighting responses. These differences can lead to systematic deviations in the results provided by different firms. To adjust for these effects, the model includes terms that account for the biases introduced by individual polling organizations, ensuring that the estimates are not skewed by the peculiarities of any single pollster.

Population Effects: Population effects stem from the differences in the demographic and political composition of the populations that different polls sample. Polls may target different subsets of the population, such as registered voters, likely voters, or the general adult population, each of which might have different voting intentions. The model incorporates adjustments to account for these differences, ensuring that the estimates reflect the overall voting population rather than any specific subgroup.

Mode Effects: Mode effects arise from the different modes of data collection used in polling, such as telephone interviews, online surveys, or face-to-face interviews. Each mode has its own potential biases. For example, telephone surveys might miss younger voters who are less likely to have landlines, while online surveys might over-represent tech-savvy individuals. The model includes adjustments for these mode effects to ensure that the polling data is comparable across different collection methods.

$$\mu^c \sim \text{Normal}(0, \sigma^c) \tag{6}$$

$$\mu^r \sim \text{Normal}(0, \sigma^r) \tag{7}$$

$$\mu^m \sim \text{Normal}(0, \sigma^m) \tag{8}$$

2.5.1.3 Partison nonresponse adjustment term   The authors suggest that nonresponse bias occurs when certain groups, such as partisans, are less likely to respond to polls, leading to systematic errors in estimating voter support.

To mitigate this issue, the model incorporates an autoregressive process to allow the partisan adjustment to vary over time at the national level. This process helps account for the dynamic nature of partisan nonresponse, which can fluctuate throughout the election cycle. Specifically, the adjustment term is designed to capture the difference between polls that adjust for partisanship and those that do not.

$$\epsilon_1|\rho, \sigma^\epsilon \sim \text{Normal}\left(0, \frac{1}{\sqrt{1-\rho^2}}\sigma^\epsilon\right) \tag{9}$$

$$\epsilon_t|\epsilon_{t-1}, \rho, \sigma^\epsilon \sim \text{Normal}\left(\rho\epsilon_{t-1}, \sigma^\epsilon\right), \quad \text{for } t = 2, \dots, T \tag{10}$$

$$\rho \sim \text{Normal}(0.7, 0.1) \tag{11}$$

This adjustment helps to estimate the extent to which non-adjusting polls are biased, ensuring a more accurate aggregation of polling data by accounting for the partisan composition of the sample. This correction is crucial for providing reliable election forecasts, particularly in regions where polling practices may vary significantly.

2.5.1.4 Measurament error   These terms are unidentifiable, meaning that for a given number N of polls, there are N independently and identically distributed terms. The model adjusts for uncertainty in the poll estimates to account for poll-specific design effects or deviations from truly random samples.

$$\zeta_i^{\text{national}} \sim \text{Normal}(0, \sigma^{\text{national}}) \tag{12}$$

$$\zeta_i^{\text{state}} \sim \text{Normal}(0, \sigma^{\text{state}}) \tag{13}$$

These adjustments allow the model to handle measurement errors that go beyond the stated margin of error provided by the polls, ensuring more accurate and reliable estimates of voter preferences.

2.5.1.5 Correlated state errors   the paper discusses the inclusion of state and national-level polling error terms to account for unmodeled measurement error beyond the stated margin of error in each poll. The authors treat state-level polling error terms as correlated across states using a scaled version of the same correlation matrix employed for changes in underlying opinions across states.

$$\xi \sim \text{MVN}(0, \Sigma^\xi) \tag{14}$$

This approach acknowledges that polling errors in different states can be correlated. The correlation matrix Σ^ξ is derived from past election results and relevant state-level predictors, such as education levels. Off-diagonal elements of the correlation matrix that are smaller than zero are set to zero, and the matrix is scaled to achieve the desired degree of day-to-day change.

This method ensures that the estimates reflect the assumption that, in the absence of specific data, similar states will exhibit similar trends, allowing for reasonable results in both national and state-level swings. The same relationship is considered for national polls, where the weighted sum of the state-level error term is included as the national-level error, with weights proportional to voter turnout in the previous presidential election.

2.5.2 General intuition

The fundamentals-based prediction serves as the prior for the state-level forecasts on Election Day. This prediction is based on historical data and various predictors, such as economic conditions and political factors. The random walk prior for state-level support (μ_t^b) is visualized as moving backward in time from Election Day to the current day of polling. This approach allows the model to dynamically update its prior for Election Day using poll-based forecasts.

The combined model updates the Election Day prior with the poll-based forecast by incorporating daily polling data. The hierarchical structure of the model, which includes multiple error terms, helps to avoid over-certainty that can result from simple poll averaging. This hierarchical approach enables the model to borrow strength across states and over time, ensuring that the estimates are robust and reflective of both state-level and national trends.

2.5.3 Model application

3 Results

3.0.1 State-level Democratic Vote Share (Prior vs. Posterior)

This plot compares the prior and posterior distributions of the Democratic vote share for each state. The prior distributions are based on historical data, while the posterior distributions are updated with current polling data. Each point represents the mean vote share, and the error bars show the 95% credible intervals.
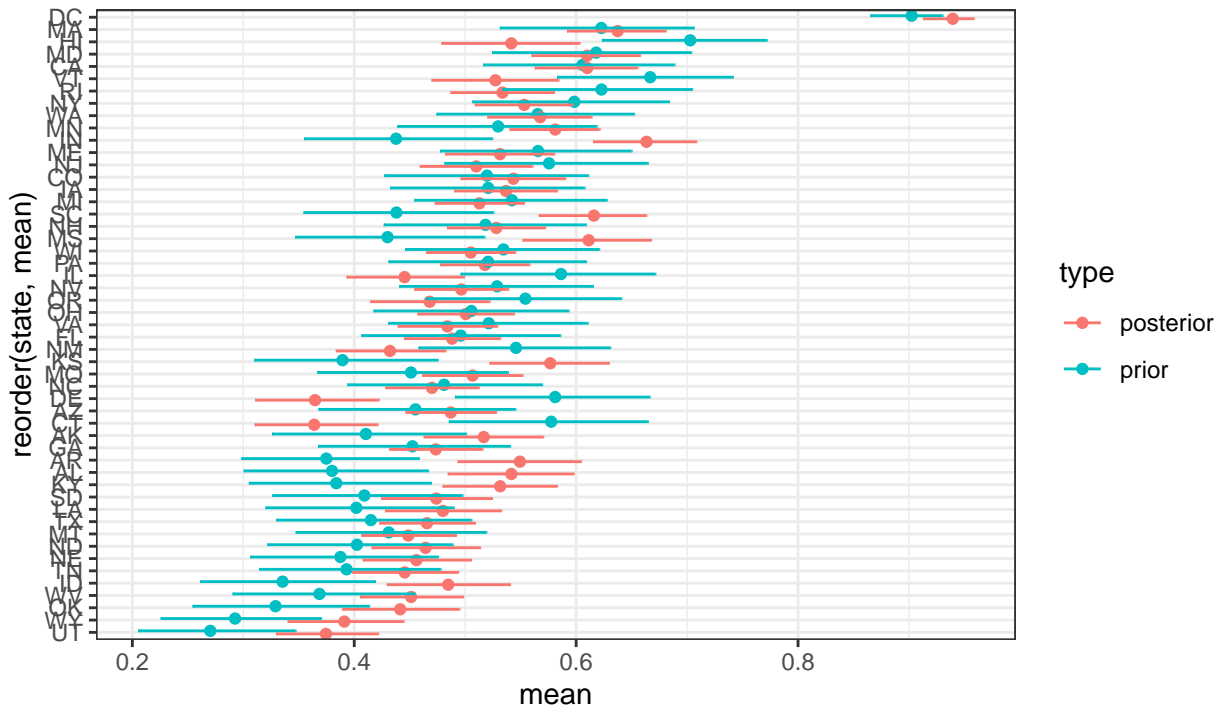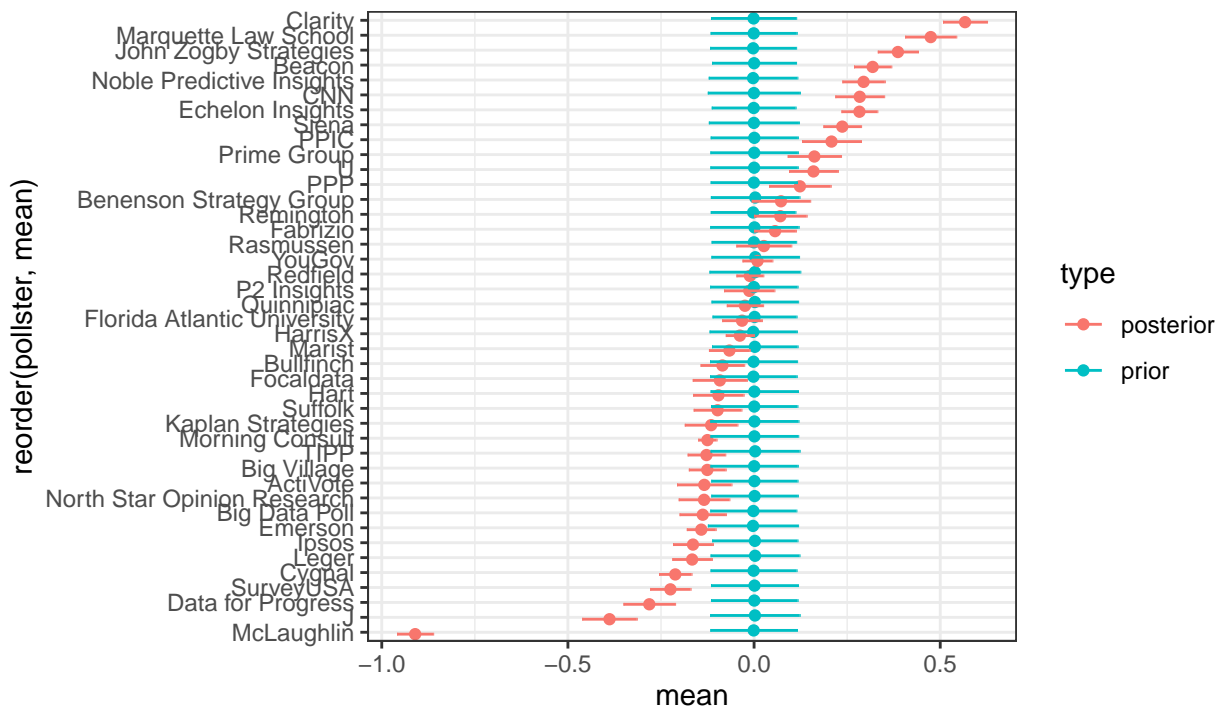

3.0.2 Pollster Effects (Prior vs. Posterior)

This plot shows the prior and posterior distributions of the pollster effects. These effects capture the systematic bias introduced by different pollsters. Each point represents the mean pollster effect, and the error bars show the 95% credible intervals.

Table 1: Summary of Data Variables

| Variable	Description	Summary
N_national_polls	Number of national polls	612
N_state_polls	Number of state polls	580
T	Total time periods	403
S	Number of states	51
P	Number of pollsters	134
M	Number of modes	44
Pop	Population vector	Mean: 4 SD: NA
state_weights	State weights vector	Mean: 0.02 SD: 0.02
day_state	Days for state polls	Min: 1 Max: 282 Median: 172
day_national	Days for national polls	Min: 1 Max: 282 Median: 149
poll_state	State polls data	Min: 1 Max: 134 Median: 59.5
poll_national	National polls data	Min: 3 Max: 134 Median: 68
poll_mode_national	National poll modes	Min: 1 Max: 44 Median: 33
poll_mode_state	State poll modes	Min: 1 Max: 44 Median: 25
poll_pop_national	National poll populations	Min: 1 Max: 3 Median: 2
poll_pop_state	State poll populations	Min: 1 Max: 4 Median: 1
unadjusted_national	Unadjusted national polls	Min: 0 Max: 1 Median: 1
unadjusted_state	Unadjusted state polls	Min: 0 Max: 1 Median: 1
n_democrat_national	Democrat counts in national polls	Min: 0 Max: 9592 Median: 1081
n_democrat_state	Democrat counts in state polls	Min: 0 Max: 2614 Median: 269.5
n_two_share_national	Two-party share in national polls	Min: 0 Max: 9738 Median: 2107
n_two_share_state	Two-party share in state polls	Min: 50 Max: 4444 Median: 523.5
sigma_measure_noise_national	Measurement noise for national polls	Mean: 0.04 SD: NA
sigma_measure_noise_state	Measurement noise for state polls	Mean: 0.04 SD: NA
mu_b_prior	Prior for state effects	Mean: -0.02 SD: 0.52
sigma_c	Standard deviation of common error	0.06
sigma_m	Standard deviation of mode effect	0.04
sigma_pop	Standard deviation of population effect	0.04
sigma_e_bias	Standard deviation of bias error	0.02
state_covariance_0	Initial state covariance matrix	Covariance matrix (too large to summarize)
polling_bias_scale	Polling bias scale	0.05
mu_b_T_scale	Scale for prior election day prediction	0.13
random_walk_scale	Scale for state level random walk	0.01

3.0.3 Mode Effects (Prior vs. Posterior)

This plot displays the prior and posterior distributions of the mode effects. Mode effects capture the biases introduced by different polling methods (e.g., phone vs. online surveys). Each point represents the mean mode effect, and the error bars show the 95% credible intervals.
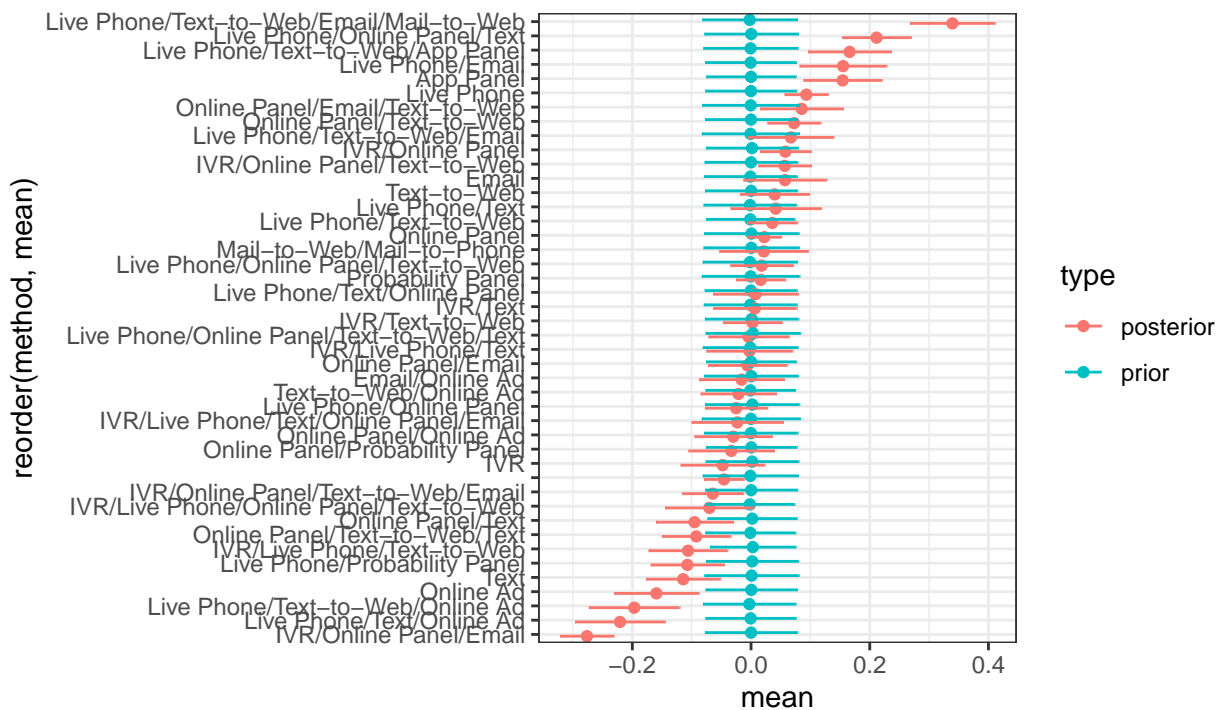
3.0.4 State-level Polling Bias (Prior vs. Posterior)

This plot shows the prior and posterior distributions of the state-level polling biases. Polling bias reflects the systematic errors in polling data for each state. Each point represents the mean bias, and the error bars show the 95% credible intervals.
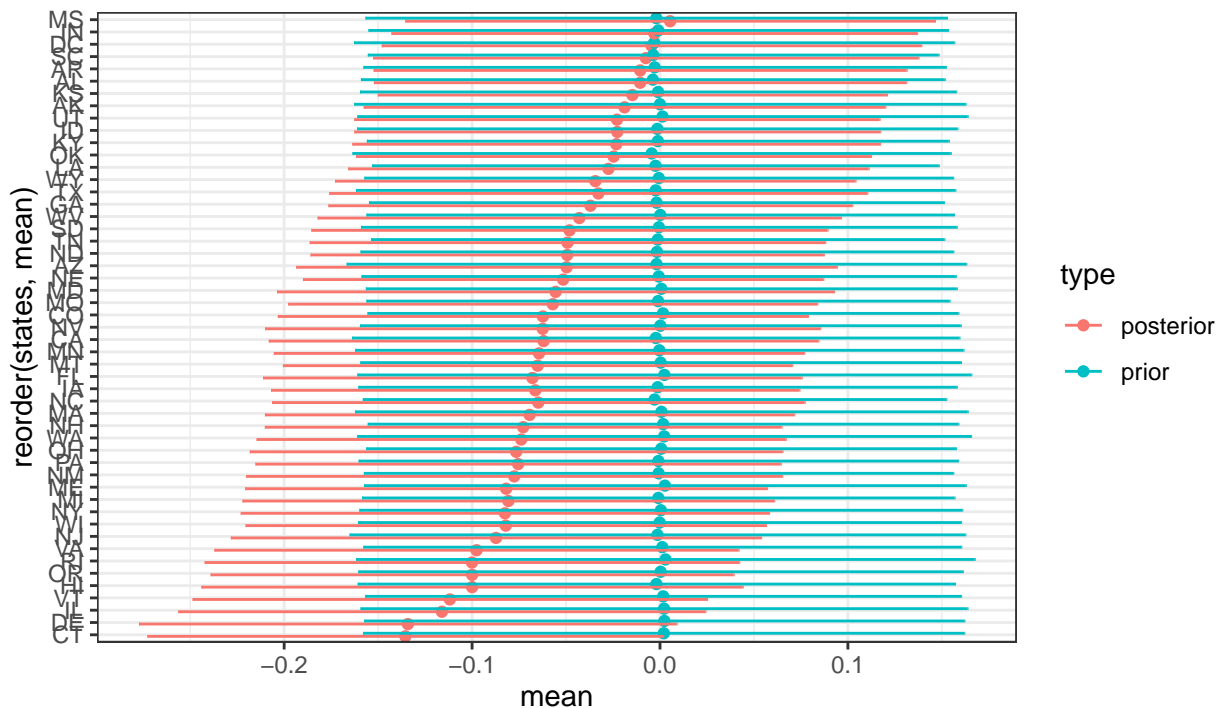

3.0.5 Adjusted Polling Error Bias Over Time

This plot tracks the adjusted polling error bias over time. The adjusted bias is computed by correcting for the mean biases of non-adjusted pollsters. Each line represents a different simulation trial.
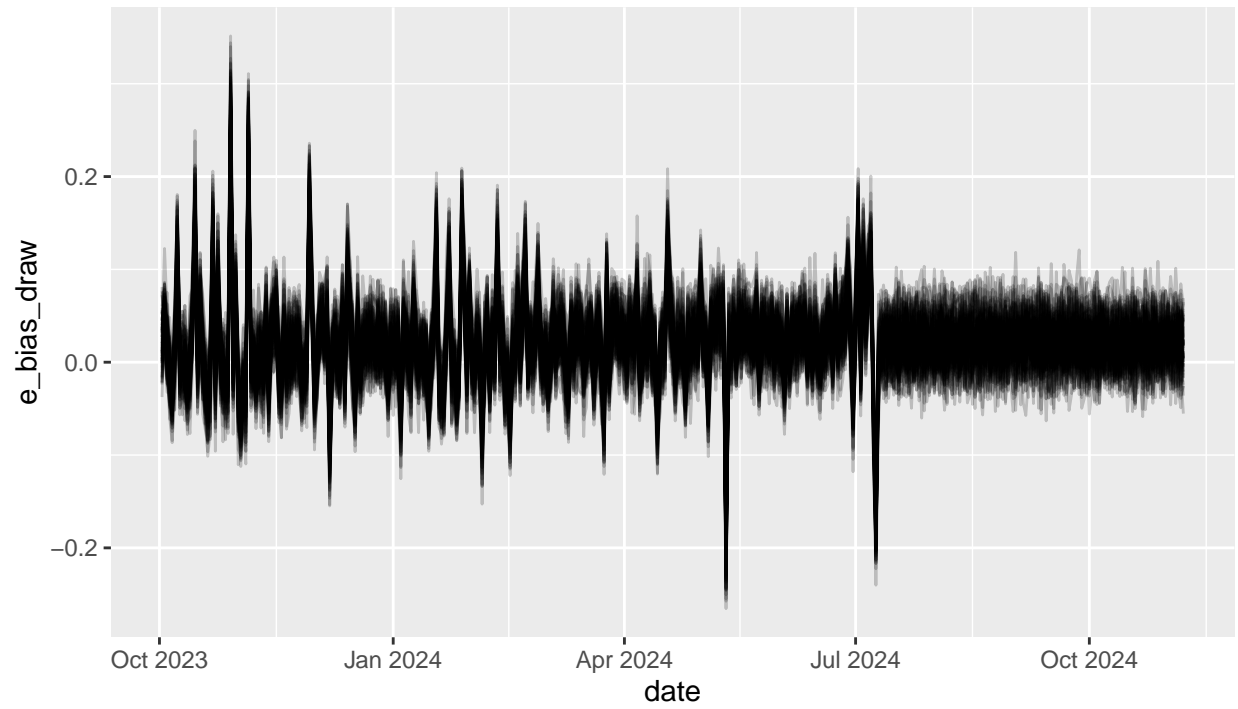
3.0.6 Unadjusted Polling Error Bias Over Time

This plot shows the unadjusted polling error bias over time. The unadjusted bias includes the raw errors from all pollsters. Each line represents a different simulation trial.
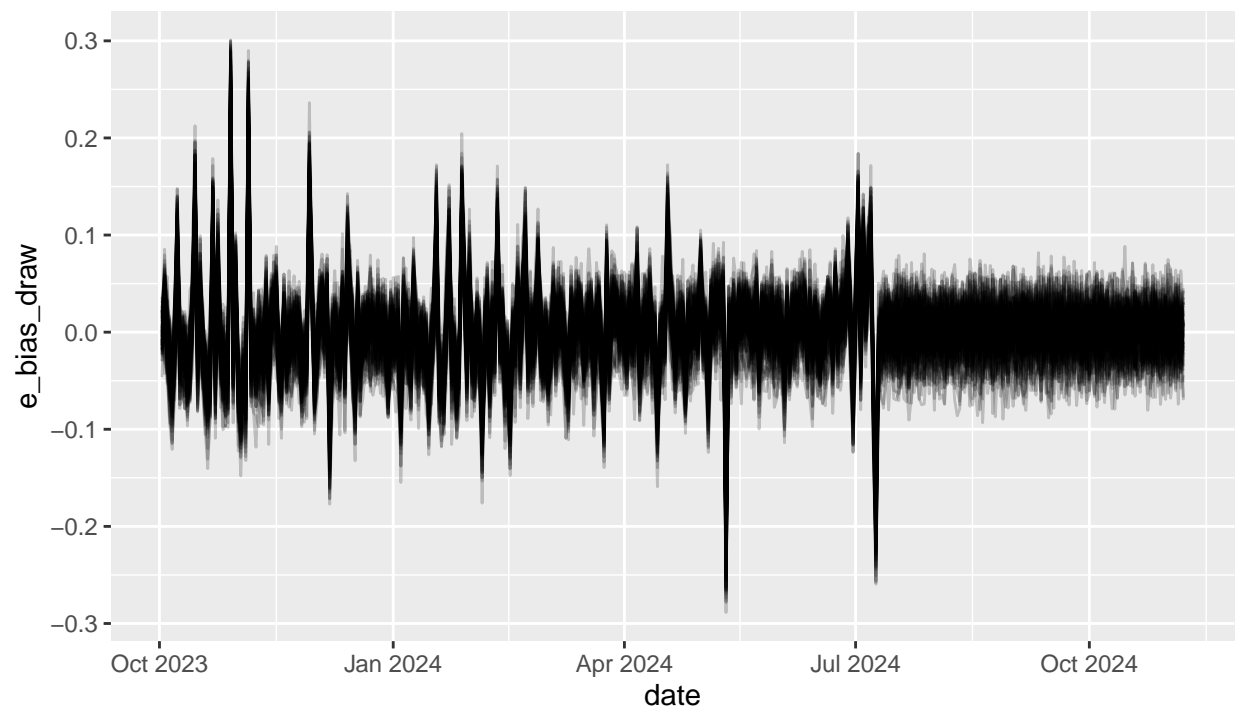
3.0.7 Predicted Percent Vote for Biden

This plot shows the predicted percent vote for Biden across different states and time periods. It includes the 95% credible intervals and the probability of winning for each state.
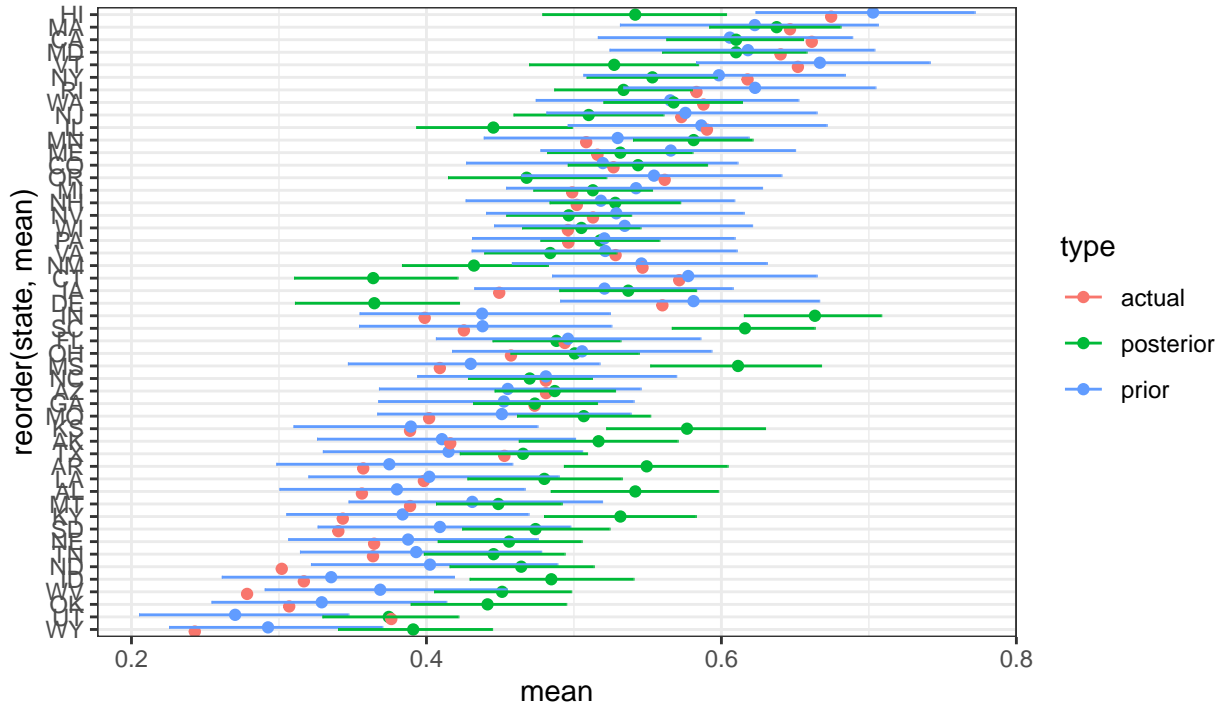
3.0.8 Probability of Biden Winning by State

This map visualizes the probability of Biden winning each state based on the current polling data. The colors range from red (low probability) to blue (high probability), with white indicating a 50% probability.
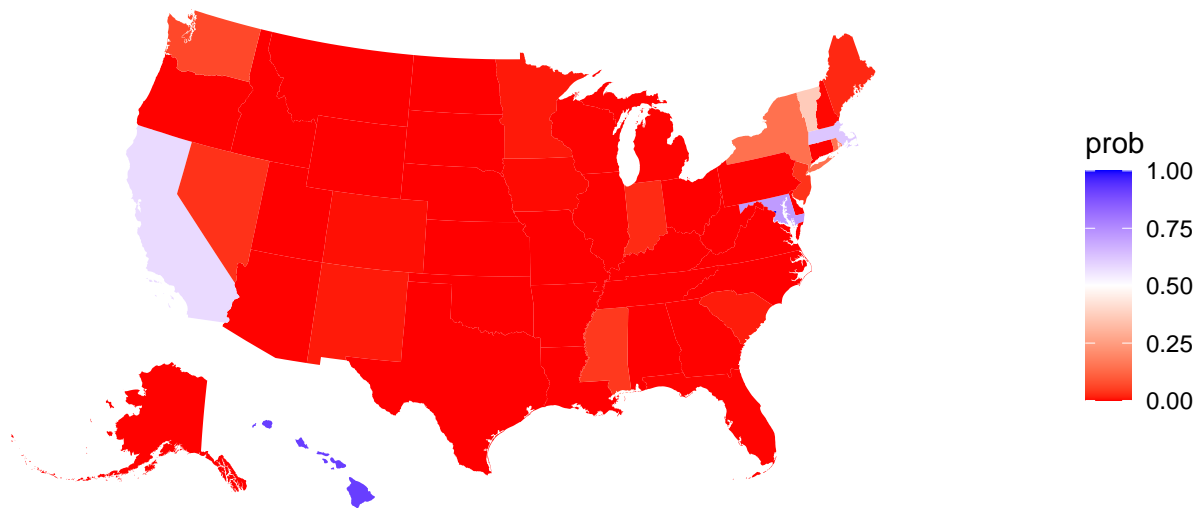
3.0.9 National Polls Over Time, Predicted Electoral Votes Over Time, and State Polls Over Time

The first plot shows the national polling data for Biden over time. It includes the mean predicted vote share, the 95% credible intervals, and individual poll results.

The second plot shows the predicted number of electoral votes for Biden over time. It includes the median number of electoral votes, the 95% credible intervals, and the probability of winning 270 or more electoral votes.

The third plot shows the state-level polling data for Biden over time for selected states. It includes the mean predicted vote share, the 95% credible intervals, and individual poll results.
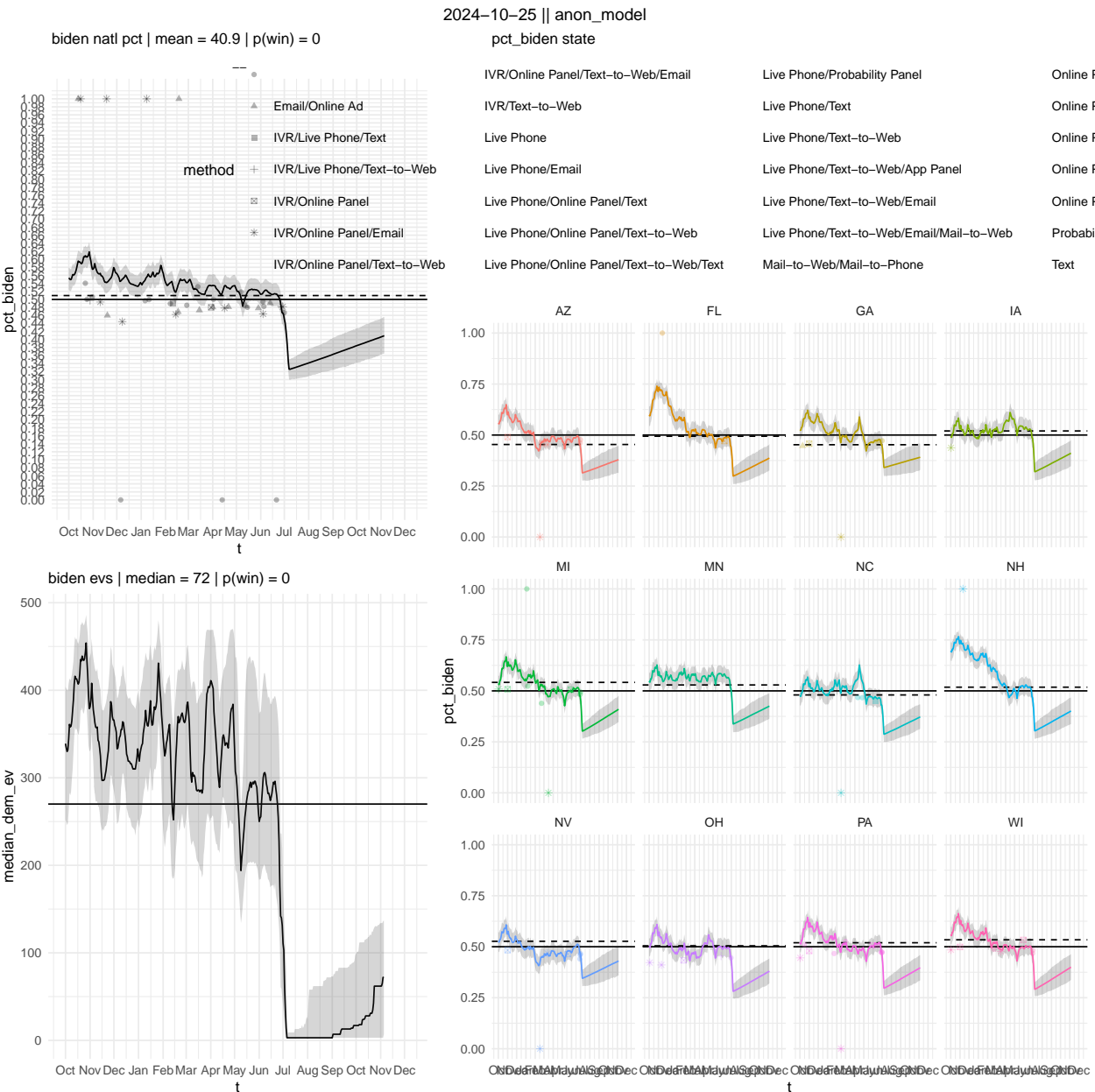
### 3.0.10 Difference from National Vote Over Time

This plot shows the difference between the predicted state-level vote share for Biden and the national vote share over time. It highlights the states where Biden's support deviates significantly from the national trend.

2024−10−25 || anon_model
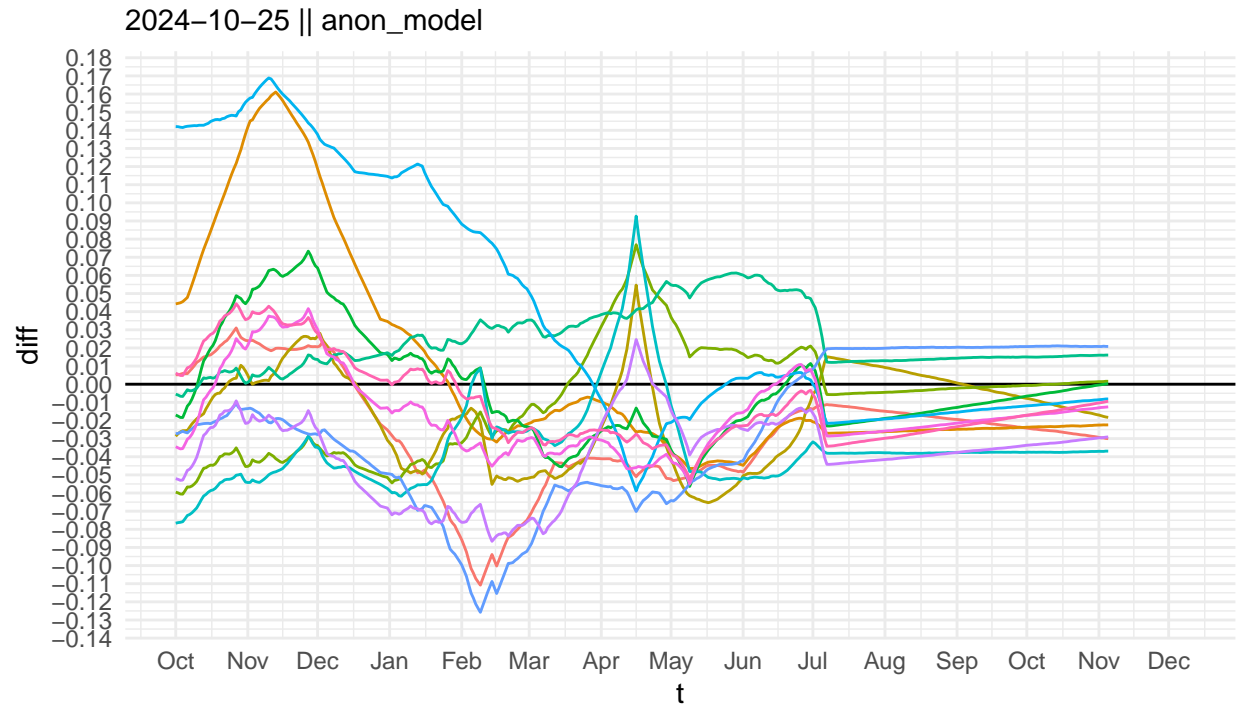
3.0.11 Final Electoral Vote Distribution

This plot shows the distribution of predicted electoral votes for Biden based on the final simulations. It highlights the probability of Biden winning the required 270 electoral votes.

p(dem win) = 0

Electoral College winner ■ Republican
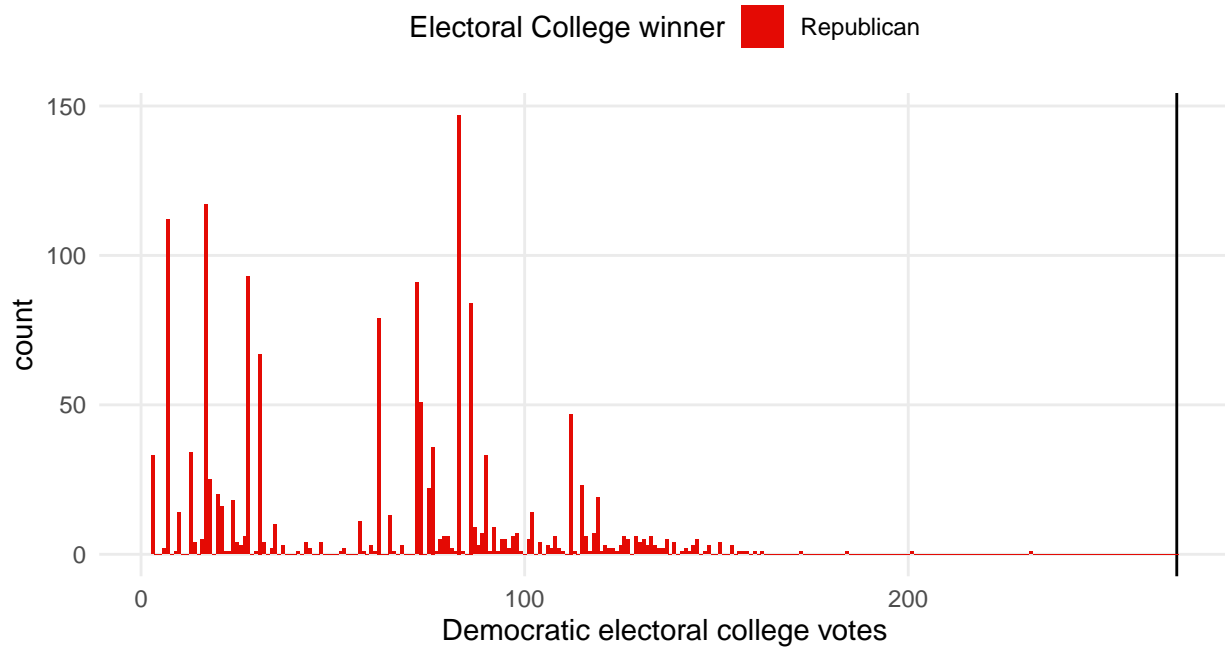
4 Conlusions

In conclusion, this dynamic Bayesian forecasting model offers a robust and sophisticated tool for election forecasting, combining rigorous statistical methodologies with practical adjustments for real-world polling biases. Key innovations in the model include the incorporation of partisan nonresponse adjustments and the careful treatment of measurement and correlated state errors. These features address some of the critical shortcomings observed in past election forecasts, such as those seen in the 2016 presidential election and the Brexit referendum. By accounting for biases and uncertainties inherent in polling data, the model provides a nuanced and reliable prediction of election outcomes. # Citations (PROVVISORIO) (Linzer 2013) (Heidemanns, Gelman, and Morris 2020) (Alexander 2024) (Rigdon et al. 2009) (Jensen, Lacombe, and McIntyre 2013) (Gelman 2024) (Stoetzer et al. 2018)

References

Abramowitz, Alan I. 2008. "Forecasting the 2008 Presidential Election with the Time-for-Change Model." *PS: Political Science & Politics* 41 (4): 691–95. https://doi.org/10.1017/S1049096508081249.

Alexander, Brittany. 2024. "A Bayesian Model for the Prediction of United States Presidential Elections."

Gelman, Andrew. 2024. "Bayesian of Election Surveys and Forecasts: Learning from the State of Valencia Experience."

Heidemanns, Merlin, Andrew Gelman, and G. Elliott Morris. 2020. "An Updated Dynamic Bayesian Forecasting Model for the US Presidential Election." *Harvard Data Science Review* 58 (4): 123–45.

Jensen, Christa D., Donald J. Lacombe, and Stuart G. McIntyre. 2013. "A Bayesian Spatial Econometric Analysis of the 2010 UK General Election." *Papers in Regional Science* 92 (3): 651–67. https://doi.org/10.1111/j.1435-5957.2012.00415.x.

Linzer, Drew A. 2013. "Dynamic Bayesian Forecasting of Presidential Elections in the States." *Journal of the American Statistical Association* 108 (501): 124–34. https://doi.org/10.1080/01621459.2012.737735.

Rigdon, Steven E., Sheldon H. Jacobson, Wendy K. Tam Cho, Edward C. Sewell, and Christopher J. Rigdon. 2009. "A Bayesian Prediction Model for the u.s. Presidential Election." *American Politics Research* 37

(4): 700–724.

Stoetzer, Lukas F., Marcel Neunhoeffer, Thomas Gschwend, and Simon Munzert. 2018. "Forecasting Elections in Multi-Party Systems: A Bayesian Approach Combining Polls and Fundamentals."