Electronic Supplemental Material: Methods Appendix

Data

This research draws from a dataset of 15 nationally representative climate change opinion surveys conducted between 2008 and early 2016 for the Yale Program on Climate Change Communication (YPCCC) and George Mason Center for Climate Change Communication (combined n = 18,158). These surveys were conducted by GfK Knowledge Networks using probability-based online samples. Survey questions are provided below, along with question coding details.

We geolocate respondents on the basis of their ZIP+4 codes or through jittered geocoded addresses (150 m radius) provided by the survey contractors. American Community Survey (ACS) estimates of demographic and housing characteristics (Series DP05), economic data (Series DP03), and household and family data (Series S1101), were also compiled at state and congressional district levels. State- and congressional district-level data representing 2012 Presidential Democratic vote share and data on per capita CO2 emissions at the state and county level from the Fossil Fuel Data Assimilation System (FFDAS) dataset were also merged with survey responses.

In recent years, a variety of political consulting groups have assembled large-n US voter files. These files amass political, economic, demographic and consumer data for US individuals through compilation of publicly and commercially available datasets. In July 2016, YPCCC hired Catalist LLC, one of the major political consultants working in this space, to prepare a series of custom congressional-district cross-tabs using their proprietary US population-level data file. These custom crosstabs provided race by gender by age population counts of Republicans within each congressional district. For gender, Catalist reported crosstabs for "Male", "Female", and "Unknown." For race, we collapsed Catalist race categories into "White, non-Hispanic", "African-American, non-Hispanic", "Hispanic", "Other, non-Hispanic", and "Unknown." For age, we collapsed individuals into four categories "18-24 years", "25-44 years", "45-64 years", and "65 years and older."

These crosstabs draw from two individual-level variables in the Catalist voter file. For the subset of states who report party id during voter registration (n=32), we have direct population counts of Republicans and Democrats for each state and/or congressional district. However, direct counts are not available for the full set of US states. To address this missing data issue, Catalist also estimates a partisanship model for every registered voter using a variety of covariates. This model takes a value of 0 for definite Republicans and a value of 100 for def-
inite Democrats. We also requested a population count cross-tab at state and district-scales grouped by the number of individuals (e.g. presumed Republicans/Democrats) in each quintile of Catalist’s partisanship model score. This allowed us to construct geography-level cross-tabs of individuals who scored in each quantile of the Catalist partisanship model (e.g. between 0 and 20 on the Catalist partisanship model). In this way, we identified a pool of likely Republicans and Democrats across all 50 states and their districts using a standard methodology. Specifically, we treat individuals with a Catalist score between 0 and 40 as a likely Republican, and individuals with a Catalist score between 60 and 100 as a likely Democrat. Below, we report comparisons of partisan belief estimates in states where we have direct voter registration data with models built from Catalist-produced partisanship model scores for those same states. This provides empirical validation of the coding strategy.

Previously, in May of 2014, YPCCC hired Catalist to prepare a "data append" for existing YPCCC climate change surveys. This data append used GfK survey respondent data to “append” each survey respondent’s Catalist voter file information to YPCCC’s existing GfK surveys responses. YPCCC subsequently purchased a data append from Catalist for climate survey waves between late 2014 and 2016. Notably, the Catalist data append includes Catalist predicted partisanship scores for each survey respondent, as well as all demographic information Catalist possesses about those individuals. Matching between our GfK survey respondents and the Catalist append file was done by GfK in partnership with Catalist, and reported to YPCCC using anonymized respondent ids. No identifying information was shared with YPCCC. In total, the YPCCC dataset includes 5696 people who score between 0 and 40 on the Catalist partisanship score, and 6383 individuals who score between 0 and 60 on the Catalist partisanship score. Our dataset also includes 2676 people who are identified by Catalist as registered Republicans in party-id reporting states, and 2807 people identified as registered Democrats in party-id reporting states.

Importantly, the Catalist crosstabs are not simply used to estimate the share of Republicans and Democrats in each state or congressional district, information that could alternatively be extracted from CCES or other public surveys. Instead, MRP models require cross-tabs that offer population counts for each demographic subgroup in each geography. This means we need accurate estimates of the distribution of partisan types in each geography across 60 demographic categories. Correspondingly, since we rely on Catalist data for post-stratification cross-tabs, we require the Catalist data append to ensure that we model climate and energy opinions as a function of the same demographic categories that later form the basis for post-stratification.
YPCCC also has direct information about respondent demographics and partisanship that are considerably higher in quality than individual-level Catalist data. For instance, race/ethnicity is known for all respondents, and gender is not modeled - as is the case with Catalist. However, since our crosstabs are based on summations over Catalist’s voter file, it is imperative to ground our modeling in similarly structured survey-specific data.

Model

Multilevel regression and poststratification (MRP) approaches are increasingly used to model spatial distributions of public opinion (for more detailed treatments, see Park et al., 2006; Lax and Phillips, 2009; Warshaw and Rodden, 2012; Buttice and Highton, 2013). An MRP analysis involves two steps. First, individual survey responses are modeled as a function of both individual-level demographics and geography-level covariates (the "multi-level regression model"). Second, using a fitted version of this regression model, population-weighted opinion estimates for demographic-geographic subtypes are aggregated based on the subtype population distribution within each geographic subunit ("post-stratification").

As our multi-level regression model, we use a hierarchical model to estimate the relationship of individual and geography-level covariates with specific climate and energy opinions, \( h \), for a given partisan individual \( i \), represented by \( y_{hi} \). We run models for Democrats and Republicans separately. For clarity, we present the model for a single opinion only, thus dropping the indexing over \( h \). At the individual-level, we thus have:

\[
Pr(y_i = 1) = \logit^{-1}(\gamma_0 + \alpha_{race}^{i} + \alpha_{age}^{i} + \alpha_{gender}^{i} + \alpha_{time}^{i} + \alpha_{geography}^{i})
\]

where

\[
\alpha_{race}^{i} \sim N(0, \sigma_{race}^2), \text{ for } j = 1, ..., 5
\]

\[
\alpha_{age}^{i} \sim N(0, \sigma_{age}^2), \text{ for } j = 1, ..., 5
\]

\[
\alpha_{gender}^{i} \sim N(0, \sigma_{gender}^2), \text{ for } l = 1, 2, 3
\]

\[
\alpha_{time}^{i} \sim N(0, \sigma_{time}^2), \text{ for } n = 1, ..., 5
\]

Each variable is indexed over individual \( i \) and over response categories \( j, k, l, \text{and } m \) for race, age, gender, and time respectively. The geography variable is flexible, with \( g \) indexing either states (s) or counties (cd), depending on the level of geographic subunit being modeled. Time
captures the year in which respondents were surveyed.

For state models, we model the geography-level term as:

\[ \alpha_{state} \sim N(\alpha_{region} + \gamma^{drive} \cdot drive_s + \gamma^{samesex} \cdot samesex_s + \gamma^{carbon} \cdot carbon_s + \gamma^{pres} \cdot pres_s, \sigma^2_{state}), \text{ for } s = 1, \ldots, 51 \]

where \( drive \) describes the percentage of individuals who drive alone in a given state, \( samesex \) describes the percentage of same-sex households in a given state, \( carbon \) describes the level of point source carbon dioxide emissions in a given state, and \( pres \) describes the 2012 Democratic Presidential vote share in a given state.

The region variable describes the census region in which a respondent resides, the effect which is in turn modelled by:

\[ \alpha_{region} \sim N(0, \sigma^2_{region}), \text{ for } j = 1, \ldots, 9 \]

Congressional-District models have identical specifications, using boundaries associated with the 113th Congressional District maps (boundaries which will apply until redistricting after the 2020 Census) with some modifications to account for the different nested nature of geographic subunits (e.g. congressional district models also include a state random effect). For all models, we use the GLMER function in the lme4 package to estimate the model (Bates et al., 2012). To estimate model uncertainty, we construct prediction intervals by drawing from a sampling distribution for the model’s random and the fixed effects and then estimating the fitted value across the distribution.

Our model combines geographic covariates that have been shown to have broad predictive power in other studies (e.g. Lax and Phillips, 2009; Warshaw and Rodden, 2012 use percentage same-sex households as an effective proxy for liberalism) with custom variables that are strongly associated with climate and energy beliefs and behaviors (e.g. driving behavior and carbon emissions). Our model follows the validated approach described by Howe et al. (2015) and Mildenberger et al. (2016) in previous local area estimation of climate and energy opinions among the general population. Other research has used MRP models to study environmental opinion at state levels including Fowler (2016) and Eun Kim and Urpelainen (2017).

For our post-stratification stage, we use our multilevel regression model results to estimate the average opinion of each demographic-geographic individual type; for instance, we estimate the average opinion of a Latina Republican woman older than 65 living in Arizona’s 5th congressional
district. (Our cross-tabs suggest that there are 525 such individuals in AZ-5). Our model allows for 60 unique sex-race-age categories, which are then interacted with either 51 states (including D.C.) or 435 districts, generating 3,060 unique population types for the state-level model and 26,100 unique population types for district-level models. Our Catalist-derived population counts tables provide the count of each population type in each subunit for each party. Final MRP estimates weight the model estimated belief of each population type by the true population count of that type in a given geographic subunit. Let \( \vartheta_w \) describe the estimated opinion of each unique demographic-geography type, indexed over cell \( w \), and \( N_w \) give the population count for that cell, then our MRP estimate of opinion in any given geographic subunit is the weighted sum of these estimates and population counts, over state or county variable \( g \):

\[
y_{\text{mrp \ state}(cd)} = \frac{\sum_{c \in g} N_w \vartheta_w}{\sum_{c \in g} N_w}
\]

**Estimation and Validation**

We present our estimation technique here using our Republican data as an illustration. However, our approach to estimate Democrats proceeds similarly.

First, we begin by separately estimating an MRP model for each geography for 1) Republican individuals with known party id (the “party registration” model); 2) individuals who score 0 to 20 on the Catalist partisanship score model; and 3) individuals who score 0 to 40 on the Catalist partisanship score model (the “Catalist score” models).

We validate our Republican MRP results for individuals with known party id using an internal cross-validation technique and an external validation against an independent dataset, the Cooperative Congressional Election Study. For our internal cross-validation, we compare our MRP estimates to estimates derived from disaggregation, following a technique developed by Pacheco (2011). Using repeated simulations, subsamples of varying sizes were randomly selected from a large-population state and used to simulate the samples of smaller provinces or districts. We undertake this analysis using the party registration model output. At the state level, this procedure operates as follows:

- We draw 99 random samples of size \( n \) from the state with the greatest number of respondents (California). We add this sampled data to all non-Californian data in our dataset. In effect, we are simulating California as if it were a small state in our dataset with a limited number of direct observations. What we want to do is compare model performance given this limited amount of revealed data about Californians with the actual mean Californian
support observed across our entire California dataset. We call this dataset that combines non-Californian data with a Californian random sample of size n the “training set”.

• We run an MRP model using this training dataset, and use this model to predict support for a given opinion among Californian Republicans. We compare the predicted opinion from the training set with the observed mean of the random sample of Californian Republicans of size n. We calculate the mean absolute error of the MRP prediction against the simple sample average.

• We repeat steps 1 and 2 for the 25th, 50th, 75th and 90th percentile of state sample sizes (n = 7, 20, 63, 133). We also track the mean absolute error as compared to the average of sampled Californian Republicans

• We repeat steps 1 through 3 for Florida, and for two sample questions.

For the disaggregation method, we estimate the opinion levels for each state using only the data from respondents in the state. Obviously, estimates using the disaggregation method are very unreliable for states with few respondents. For states with large number of respondents, such as California and Florida, the disaggregated estimates are relatively close to the MRP estimates. More importantly, the MRP at low state-level sample sizes quickly converges to the disaggregation estimates for large-sample sizes (see Fig 6).

Our crossvalidation method differs slightly at the congressional district level to account for a roughly uniform distribution of respondents per district (Figure 7). Since districts are close in population, we lack districts with a large number of respondents to serve as our "baseline" for validation, as with California and Florida in the state-level model. Instead, we construct simulated "megadistricts," groups of more than one district, to serve as our validation baseline. Megadistricts were constructed by dividing all districts into terciles based on their estimated belief that global warming is happening, randomly sampling 15 districts within each tercile, and aggregating sampled districts together to form a larger simulated district baseline against which to measure MRP estimates. Each megadistrict contains between 150 and 200 respondents. Similarly to the state-level validation, we construct 4 simulated district sample sizes based on the 25th, 50th, 75th and 90th percentile of district sample sizes (n = 3, 8, 12, 16), repeating the steps for each of the three megadistricts. At the district level the MRP model outperforms disaggregation across all sample sizes. The absolute errors of the MRP estimates are largely consistent across simulated district sample sizes, likely because of the small variation in actual district sample sizes in our dataset.
Figure 6: Cross-validation comparison of mean absolute error between MRP results and disaggregation against the full sample, across 4 simulated sample sizes (n = 99 simulations) for California and Florida. The analysis is for two variables: opinion that global warming is happening (top) and support for renewable portfolio standards (RPS) policies (bottom).

Next, our external validation compares our modeled MRP estimates to an independent dataset, disaggregated survey responses from the Cooperative Congressional Election Study (CCES) in 2016. We compare our estimates in party registration states for the support for carbon regulation item to state-level data from the CCES for a similar question: "Do you support or oppose each of the following proposals? [Give Environmental Protection Agency power to regulate Carbon Dioxide emissions]." This item is differently worded from our survey item and as such we would not expect absolute support to match between the two items. However, we would reasonably expect support for each item to be correlated across subnational geographies.
Figure 7: Cross-validation comparison of mean absolute error between MRP results and disaggregation against the full sample across 4 simulated sample sizes (n = 99 simulations) for 3 simulated "megadistricts." Megadistricts constructed by dividing all districts into terciles based on their estimated opinion that global warming is happening, randomly sampling 15 districts within each tercile, and aggregating sampled districts together to form a larger simulated district baseline with which to measure MRP estimates. The analysis is for two variables: opinion that global warming is happening (top) and support for renewable portfolio standards (RPS) policies (bottom).

We select respondents who identify as registered Republicans in the 14 states with at least 200 responses in the CCES dataset (n=8,466). Our MRP estimates correlate with the disaggregated CCES state-level data at r=0.75. Figure 8 plots the CCES disaggregated estimates for these 14 states against our MRP estimates. As the figure illustrates, the model estimates exhibit lower variance as compared with the external survey data, which is a common pattern in model-based small area estimation. By contrast, direct survey measurements exhibit the opposite pattern.
Figure 8: External validation comparison of support for carbon regulation between CCES disaggregated estimates and MRP estimates for 14 states with party registration and greater than 200 respondents in the CCES dataset. Axes represent differences from the national mean to control for question wording effects.

and are more likely to have greater variance due to survey measurement error.

We then move on to evaluating congruence between estimates derived from the party registration model (n=32 states) and estimates from the Catalyst score models (for which data is available for all 50 states). Ideally, we would like to see strong correlation between estimates derived using these two separate data sources. (However, we should tolerate mean shifts in the estimates as the relevant population being modeled is slightly different: the pool of individuals who score between 0 and 40 on the Catalyst partisanship model score may include Republican-leaning independents who are not registered Republicans in their states. Likewise the pool of
individuals who score between 0 and 20 on the Catalist partisanship model score may include registered Republicans who otherwise profile as weaker partisans.)

Initial analysis suggested that the Catalist 0 to 20 score model underestimated opinion relative to direct party registration model, while the Catalist 0 to 40 score model overestimated opinion relative to direct party registration model. However, a simple average of the 0 to 20 and 0 to 40 estimates closely matches the direct party registration model. We thus blended the 0 to 20 and 0 to 40 models into a single Catalist score model estimate. Figure 9 charts the correlation between party registration model estimates and predictions generated by Catalist score model for states with estimates from both models, for all questions reported in this research letter. Figure 10 provides the same results but for the congressional district level instead. All models have an r value above 0.97, except the state-level correlation for human-caused climate change which has an r value of 0.90.

For states that are part of the party registration model, we present model results as our estimates. For states without party registration data (n=18), and all congressional districts in those states, we interpolate results using Catalist score model output. More precisely, we predict opinion in these states and congressional districts from a linear regression of the party registration model on the Catalist score model.4

In Figures 11 and 12 we graphically present analogous figures for our estimates of Democratic climate and energy opinions. Again, we find that averaging our models that draw from Catalist party id scores accurately recovers estimates independently produced directly from party registration cross-tabs. Just as with our Republican estimates, for states that are part of the party registration model, we present model results as our estimates. For states without party registration data (n=18), and all congressional districts in those states, we interpolate results using Catalist score model output.

4In other word, we model the relationship between output from these two distinct models for states that appear in both (n=32), and predict values for states with missing party registration data using those states’ Catalist score model estimates.
Figure 9: Comparison of Republican estimates derived using direct party registration cross-tabs and Catalist partisanship model score cross-tabs, for all states with available party registration cross-tabs
Figure 10: Comparison of Republican estimates derived using direct party registration cross-tabs and Catalist partisanship model score cross-tabs, for all congressional districts in states with available party registration cross-tabs.
Figure 11: Comparison of Democratic estimates derived using direct party registration cross-tabs and Catalist partisanship model score cross-tabs, for all states with available party registration cross-tabs
Figure 12: Comparison of Democratic estimates derived using direct party registration cross-tabs and Catalist partisanship model score cross-tabs, for all congressional districts in states with available party registration cross-tabs
Survey Questions

Here, we provide the specific question wordings for questions discussed in this article, along with our formal coding strategy for these questions.

Opinion that global warming is happening

Question: Recently, you may have noticed that global warming has been getting some attention in the news. Global warming refers to the idea that the world’s average temperature has been increasing over the past 150 years, may be increasing more in the future, and that the world’s climate may change as a result. What do you think: Do you think that global warming is happening? [Response Scale: Yes, No, Don’t Know]

We code respondents who answer “yes” as 1, and all others as 0.

Opinion that global warming is human-caused

Question: Assuming global warming is happening, do you think it is... [Response scale: Caused mostly by human activities, Caused by human activities and natural changes, Caused mostly by natural changes in the environment, Neither because global warming isn’t happening]

We code respondents who answer “caused mostly by human activities” as 1, and all other responses as 0.

Opinion that there is a scientific consensus that global warming is happening

Question: Which comes closest to your own view? [Response scale: Most scientists think global warming is happening, There is a lot of disagreement among scientists about whether or not global warming is happening, Most scientists think global warming is not happening, Don’t know enough to say]

We code respondents who answer “most scientists think global warming is happening” as 1, and all other responses as 0.

Support for carbon regulation

Question: How much do you support or oppose the following policies? Regulate carbon dioxide (the primary greenhouse gas) as a pollutant. [Response scale: Strongly oppose; Somewhat oppose; Somewhat support; Strongly support]
We code respondents who answer “strongly support" or “somewhat support" as 1, and all other responses as 0.

Worried about global warming

Question: How worried are you about global warming? [Response scale: Very worried; Somewhat worried; Not very worried; Not at all worried]

We code respondents who answer “Very worried” or “somewhat worried” as 1 and all other responses as 0.

Fund research into renewable energy sources

Question: How much do you support or oppose the following policies? Fund more research into renewable energy sources, such as solar and wind power. [Response scale: Strongly support; Somewhat support; Somewhat oppose; Strongly oppose]

We code respondents who answer “Strongly support” or “Somewhat support” as a 1, and all other responses as 0.

Regulate CO2 as a pollutant How much do you support or oppose the following policies?

Regulate carbon dioxide (the primary greenhouse gas) as a pollutant

Require utilities to produce 20% electricity from renewable sources

Question: How much do you support or oppose the following policies? Require electric utilities to produce at least 20% of their electricity from wind, solar, or other renewable energy sources, even if it costs the average household an extra $100 a year [Response scale: Strongly support; Somewhat support; Somewhat oppose; Strongly oppose]

We code respondents who answer “Strongly support” or “Somewhat support” as a 1, and all other responses as 0.