Temperature and Temperament: Evidence from Twitter

Patrick Baylis*

February 22, 2019

Abstract

Can social media reveal preferences for environmental amenities? To better understand preferences for temperature, this paper demonstrates a new approach to nonmarket valuation by combining more than a billion Twitter updates with natural language processing algorithms to construct a rich panel dataset of expressed sentiment. Using multiple measures, I find consistent and statistically significant declines in expressed sentiment from both hot and cold temperatures. I exploit the richness of the data to document heterogeneity in both regional and seasonal responses, with implications for climate adaptation. I conclude with a suite of validation exercises to enable future work using this technique.

*University of British Columbia, Vancouver School of Economics; Iona 113, 6000 Iona Dr., Vancouver, BC V6T 2E8; patrick.baylis@ubc.ca. I am grateful to Maximilian Auffhammer, Severin Borenstein, and Solomon Hsiang for their invaluable suggestions, as well as to Michael Anderson, Hunt Alcott, Judson Boomhower, Josh Blonz, Marshall Burke, Fiona Burlig, Tamma Carleton, Richard Carson, Aluma Dembo, Meredith Fowlie, Walter Graf, Sylvan Herskowitz, Nick Obradovich, Elizabeth Sadoulet, seminar participants at UC Berkeley, the AERE Annual Conference, the CU Environmental and Resource Economics Workshop, the Heartland Workshop at Illinois, and two anonymous referees. This paper was previously circulated under the title "Temperature and Temperament: Evidence from a billion tweets."

As the possibility of substantial changes to Earth's climate becomes more certain, economists have become increasingly interested in calculating the full scope of benefits and costs resulting from these changes. Acute environmental stressors like typhoons, hurricanes, and other marked changes in the external environment can have dramatic and immediate impacts on economic well-being (Hsiang and Jina 2014) while more gradual environmental changes, such as temperature increases, have more subtle but perhaps costlier long-run economic impacts (Burke, Hsiang, and Miguel 2015b). Other work has estimated the cost of changes in income, health, agriculture, civil conflict, natural disasters, and other economic outcomes (Carleton and Hsiang 2016).

A subset of the literature has examined whether increased ambient temperature will induce a significant change in the amenity value of the climate itself. Because ambient temperature is non-rival and non-excludable, there are no direct markets from which researchers might infer preferences for different climates. This is a classic problem in the environmental valuation literature (Pearce 2002), and in the case of climate most prior work has relied on hedonic valuation approaches to indirectly estimate individuals' willingness-to-pay for different climate characteristics by observing how housing prices vary with climatic conditions. These approaches generally estimate that individuals would pay between 1% and 4% of their annual incomes to avoid projected end-of-century increases in temperature (Cragg and Kahn 1997; Sinha, Caulkins, and Cropper 2018; Albouy et al. 2016). However, because the climate to date has varied relatively little across time, these values are identified using cross-sectional differences in climates.

This paper demonstrates a new, cost-effective method to estimate preferences over public goods that allows researchers to include controls for unobservables across both time and space: I construct a spatially and temporally rich dataset on daily expressed sentiment, or emotional state, and estimate the relationship between sentiment and outdoor ambient temperature. Section 1 lays the groundwork by discussing how this approach relates to existing methods designed to elicit preferences for environmental goods. Section 2 describes the construction of the dataset, which begins with a geographically and temporally dense collection of more than a billion geocoded social media updates (hereafter, "tweets") from the online social media platform Twitter. I measure the expressed sentiment of each tweet using a set of natural language processing (NLP) algorithms designed to extract sentiment, or emotional state, from unstructured text data. For computational tractability and to account for noise in the estimation of expressed sentiment, the primary analysis takes daily averages for each Core-Based Statistical Area (hereafter, CBSAs) as unit of observation, although in the appendix I also estimate a model with individual tweets as the observation to test for compositional effects. Because of the uncertainty inherent in estimating underlying emotional state from language, I compile four separate measures of sentiment using word lists constructed using previous research in NLP and, in three of four cases, specifically intended to extract sentiment from "microblogs" such as tweets.

The analysis in Section 3 then uses the geographic information attached to the tweets in my dataset to match measures of sentiment to daily weather conditions at the location of the user. The identifying assumption in the econometric model I estimate is that temperature realizations are as good as random after accounting for spatial and temporal fixed effects. Allowing temperature to enter the model flexibly, I find consistent evidence of an upside-down "U" shape: a roughly symmetric decline in sentiment away from moderate temperatures, with peak sentiment occurring roughly around 22.7 C (72.7 F). The point estimate of the difference in expressed sentiment between 20-25 C and above 40 C is a statistically significant and between 0.1 and 0.2 SD, depending on measure used. The responses of expressed sentiment to temperature are markedly similar across choice of measure, and both qualitatively and quantitatively consistent across a range of different specification choices. As a means of understanding the mechanism by which sentiment responds to temperature, I also estimate the relationship between online profanity and temperature and find a U-shaped relationship there as well, suggesting that aggression is at least part of the explanation for the decline in expressed sentiment in both hot and cold temperatures.

To better understand how preferences for climate are formed and updated, Section 4 extends the baseline results by examining seasonal and regional heterogeneity in the response of sentiment to temperature. I find notable differences in both. Seasonally, the responses suggest preferences for cooler temperatures in summer and fall and warmer temperatures in winter, with relatively little sentiment response to temperatures on the spring. Regionally, areas that are colder tend to have stronger response to warm temperatures, and vice-versa. Finally, I document temporal features in the response using a cumulative dynamic lag model, which I estimate using deviations from the contemporaneously preferred temperature observed above. While there is no statistically significant lagged impact of colder-than-preferred temperatures on expressed sentiment, the lagged impacts for warmer-than-preferred temperatures are positive, and in fact negate and then reverse the contemporaneous impact after a few days.

Section 5 documents a set of validation exercises designed to aid interpretation of the magnitude of the results in previous sections. I begin by showing how average sentiment changes over the course of the week, and that the difference in expressed sentiment between a Sunday and Monday is roughly 0.1 SD. Next, I present three distinct empirical exercises identifying the effect of plausibly random variation in hurricanes, American football outcomes, and receipt of parking or speeding tickets on observed social media sentiment. I discuss the implications of these findings for the existing literature on the value of climate amenities and assess the potential for social media data to serve as a complementary measure of nonmarket valuation.

This paper makes several contributions to the literature. It is the first to identify sentiment-analyzed social media posts as a source of information on latent individual preferences for environmental goods. I identify a response function of sentiment to temperature that concurs qualitatively with existing work and is robust to a wide range of statistical specifications, and the methods I document provide a tractable roadmap for future work. Second, the paper identifies sharply diverging seasonal and regional preferences for temperature, strongly suggesting an important adaptive component to the baseline observed response. Finally, the set of validation exercises I employ serve as novel empirical exercises on their own and provide an initial step in identifying how

1 Background

Economists have studied the economic impacts of climate change for more than two decades (Nordhaus 1991; Cline 1992), but the increasing availability of a range of panel datasets have made possible the identification of the causal effects of changes in temperature on a diverse set of economic outcomes, including crop yields, economic production, civil conflict, mortality, migration, and many others (Carleton and Hsiang 2016). In the absence of historical changes in long-run climate, researchers have used estimates of the changes in these outcomes resulting from plausibly exogenous historical variation in temperature in order to predict future damages from climate change (Dell, Jones, and Olken 2014). The assumptions required for this extrapolation are formally derived in Hsiang (2016), but intuitively the central concern is whether or not adaptation behaviors require large fixed costs. This question is likely to be answered differently for different sectors, but for those in which it is econometrically possible to estimate a "long-differences" approach, e.g., Burke and Emerick (2015), few differences have been found between estimates produced using short-run or long-run variation in temperature.

This work has had an impact on public policy. Many of the estimated outcomes contribute, directly or indirectly, to aggregate measures of the total cost of climate change produced by summary reports (Stern 2006; Houser et al. 2014) and integrated assessment models (IAMs), which in turn are inputs to the United States government's estimate of the social cost of carbon, or SCC (Interagency Working Group on Social Cost of Carbon 2013).[1] By 2014, the then-central value of \$36 per ton of CO_2 equivalent had been incorporated into 79 U.S. regulations as part of required benefit-cost analyses conducted in the course of the federal rule-making (United States Government Accountability Office 2014).[2]

Different areas of the world will experience climate change in very different ways. Coastal areas will face rising sea levels and major economic impacts from typhoons or hurricanes (Hsiang 2010). Farmers are likely to experience substantial changes in the yields of major crops (Schlenker and Roberts 2009), and many areas in the developing world where subsistence farming is a major source of calories could experience catastrophic droughts and resulting food security crises (Lobell et al. 2008). For others, the impacts of climate change will be more subtly felt: instead of increases in large-scale natural disasters or acute economic crises, most of the world will simply experience a steady increase in average temperatures (IPCC 2014). Prior work has projected the impact of these gradual changes on income (Deryugina and Hsiang 2014), crime (Ranson 2014), mortality (Deschênes and Greenstone 2011), and other outcomes. This paper focuses instead on the welfare cost of changes in amenity values resulting from rising outdoor temperatures.[3]

Traditional approaches to calculating the welfare impact of a policy change date back as far as Marshall (1890) and rely on knowledge of either the demand curve, the supply curve, or both. For private goods with well-established markets, the shapes of these curves can be estimated using plausibly exogenous supply or demand shifters and from those the change in welfare due to a change in policy can be calculated. Estimating changes in welfare due to changes in the allocation of public goods, or nonmarket goods more generally, has proven to be more challenging due to the absence of available markets. Nevertheless, a handful of approaches to this problem have emerged, many within the environmental economics literature (Pearce 2002).

Climate can be viewed as a public amenity[4]: it is non-rival (a single person's consumption of climate does not reduce the amount of climate available to anyone else) and non-excludable (no person cannot be prevented from consuming climate), and although individuals can alter their local climates at home and at work, the outdoor ambient tem-

1. Three IAMs are used to derive the social cost of carbon: DICE, FUND, and PAGE. Diaz (2014) provides a more detailed comparison of how these models are built. 2. The Environmental Protection Agency, under direction from the Trump administration, reduced the SCC to between \$1 and \$6 per ton, largely because it used only damages to the United States in its accounting for the costs of carbon emissions.
3. To date, only the DICE model directly incorporate estimates of the costs or benefits from climate as an amenity.Nordhaus and Boyer (2000) finds that 2.5 C warming will result gain of about 0.3% of GDP. 4. I use the term "public amenities" here to indicate that changes in the climate can be either goods or bads, depending on the region and sector of the world in which they occur.

perature is determined by factors outside of their control. Hedonic price approaches provide a method to value amenities like climate: recent work by Sinha, Caulkins, and Cropper (2018) and Albouy et al. (2016) identify implicit values for different climates using observed choices about household decisions on where to live. These approaches are useful in part because it is straightforward to back out monetary valuations of different climates from the model estimates. However, since historical climate changes thus far have been modest at best, the estimates from these models must be identified using cross-sectional variation. As a result, unobserved spatial variation such as cultural norms, geographic factors like proximity to oceans or mountains, or other unobserved amenities that correlate with climate could bias these estimates in unknown directions.

A related approach to understanding preferences is to use surveys of subjective well-being (SWB) to estimate preferences over temperature. These surveys ask respondents to assess their well-being on a single dimensional scale (Diener 2000; Dolan, Peasgood, and White 2008). Kahneman and Krueger (2006) and Mackerron (2012) discuss the merits and weaknesses of these studies: a common challenge is that measurements of SWB are by definition subjective and likely to include unobserved variation across time and space. For example, responses to questions about one's well-being may depend on regional dialects or norms, or could be driven by the interaction between the interviewer and the interviewee, which may itself be affected by temperature.

The estimates of the effect of temperature on SWB vary widely within the literature. Most studies use cross-sectional variation or follow a very small group of individuals over time. To my knowledge, only two control for unobservable cross-sectional variation using panel data methods. Feddersen, Metcalfe, and Wooden (2012) use nearly 100,000 observations from Australian SWB surveys to compare the effects of short-term weather and long-term climate on life satisfaction. Since individuals are observed more than once in their data, they are able to control for individual fixed effects for some specifications. They find that weather affects reported life satisfaction through solar exposure, barometric pressure, and wind speed, but do not find impacts from changes in temperature itself. Dennisenn et al. (2008) uses an online survey to find that weather impacts are variable across individuals, but that those variations do not correspond to observable characteristics.

A small literature attempts to assign a monetary value to environmental goods using self-reported happiness data (Welsch and Kühling 2009). For example, Rehdanz and Maddison (2005) estimate the relationship between climate and self-reported happiness, and include a valuation method based on country-level GDP. Levinson (2012) conducts a similar exercise to estimate WTP to avoid pollution using happiness data, but includes weather as a covariate. Because these studies implicitly rely on income as an exogenous driver of happiness, this approach could induce bias if that assumption does not hold (Mackerron 2012).

The method of assessing preferences for nonmarket goods I describe in this paper relies on the assumption that contemporaneous changes in expressed sentiment deliver insights into individuals' underlying preferences for these goods. I have described previous work that assesses these preferences and the challenges they face in controlling for unobservable sources of cross-sectional variation. The approach in this paper mitigates the problem of unobserved correlates over time and space, allows for flexible estimation of non-linear effects, and even provides sufficient data richness to examine lagged responses and geographic variation in the response of expressed sentiment.

Conceptually, one way to view expressed sentiment is as an estimate of "experienced utility". The concept of experienced utility predates the modern neoclassical definition of utility, which for clarity and followingKahneman and Sugden (2005), I refer to hereafter as "decision utility". Whereas decision utility is an ordinal description of the value obtained from bundles of goods, experienced utility follows is an instantaneous measure pleasure and pain Bentham (1789). Discussions of which measure of utility is the appropriate metric for welfare analyses is beyond the scope of this paper, but for the purposes here it is sufficient to view expressed sentiment, like subjective well-being or experienced utility, as a useful proxy for individual preferences. In the following section, I describe how this paper estimates expressed sentiment and document descriptive statistics that suggest its relationship to underlying preferences.

2 Data

While it would be prohibitively expensive to estimate daily sentiment across the United States using a survey, publicly available updates on social media provide a low-cost alternative. By combining a large set of geo-located tweets with sentiment analysis algorithms (NLP algorithms designed to elicit emotional state), I am able to measure daily variation in expressed sentiment across the United States. In this paper, I combine this data with meteorological observations to estimate the sentiment response to temperature. Previous work in computer science has estimated models that related expressed sentiment to meteorological variables (Hannak et al. 2012), but to the best of my knowledge this is the first paper to do so in a causal framework and in order to elicit underlying preferences for temperature. The following section describes the construction of the measures of sentiment and the weather covariates. Table 1 summarizes the variables included in the empirical model. The first panel shows the count, mean, median, minimum, and maximum of the measures of sentiment, the second panel describes the weather data used, and the third panel summarizes the number of tweets by CBSAs and by individuals in the data.

Table 1: Sample characteristics

	Count	Mean	Median	Min	Max
A: Sentiment measures					
AFINN-111	687,252,523	0.5	0.5	-5	4
Hedonometer	1,220,553,949	5.5	5.5	2.8	8.3
LIWC	1,258,416,865	0.3	0.2	-3	4
Vader	1,337,317,534	0.1	0.1	-1	1
B: Weather covariates					
Minimum temperature (C)	1,337,317,534	9.1	10.8	-33.7	31.5
Maximum temperature (C)	1,337,317,534	20.9	23.8	-23.9	47.2
Precipitation (mm)	1,337,317,534	3	0	0	372.7
C: Twitter updates per...					
CBSA	900	1,485,908	190,834	30	92,894,251
User	12,852,098	123	11	1	396,822

Notes: First panel summarizes unstandardized measures of expressed sentiment: AFINN-111, Hedonometer, LIWC, and Vader. Second panel summarizes weather covariates obtained from PRISM. Third panel summarizes the number of tweets per CBSA and user.

2.1 Twitter data

Created in 2006, Twitter is a social media platform where users exchange brief updates, otherwise known as tweets. Since its founding, Twitter has become one of the most popular such platforms worldwide, with 288 million active users sending over 500 million tweets per day as of 2015.[5]

Twitter's Streaming API is designed to give developers access to the massive amount of data generated on the Twitter platform in real time. Starting in June 2014, I began collecting geolocated Twitter updates from within the continental United States using a client that is continuously connected to the Streaming API.[6] I collect the vast majority of geolocated tweets produced within my sample period, which ends in October 2016.

Geo-located tweets are those for which the user has consented to have his or her location information shared. The location information is either produced using the exact latitude and longitude or from a reverse-geocoding algorithm that derives the latitude

5. Per Twitter's website, accessed September 2015.

6. There are two substantial gaps, from June 26th to July 12th, 2014, and from September 18th to October 27th, 2014, and a small number of gaps of a few days. These gaps correspond to periods of time when the streaming client was unable to connect to the Streaming API.

and longitude from location information entered by the user. In principle, Twitter limits the total number of tweets delivered through the Streaming API to 1% (Morstatter, Pfeffer, Liu, and Carley 2013) of the total tweets created. Since I request only geolocated tweets from within the United States, this total infrequently comes to more than 1% of the total tweets worldwide (geocoded and otherwise). As a result, over the course of the days in which the streaming client was operational, the percentage of missed tweets is fewer than 0.01% of the total geolocated tweets within the United States. The left panel of Figure 1 maps the total tweet volume in my sample across the United States, where pixel shading represents the logged volume of tweets. There is considerable spatial variation in Twitter activity, and most activity occurs in cities. The map also captures the extent to which this activity follows human movement patterns: along with cities, major highways and roads are readily visible in the map.

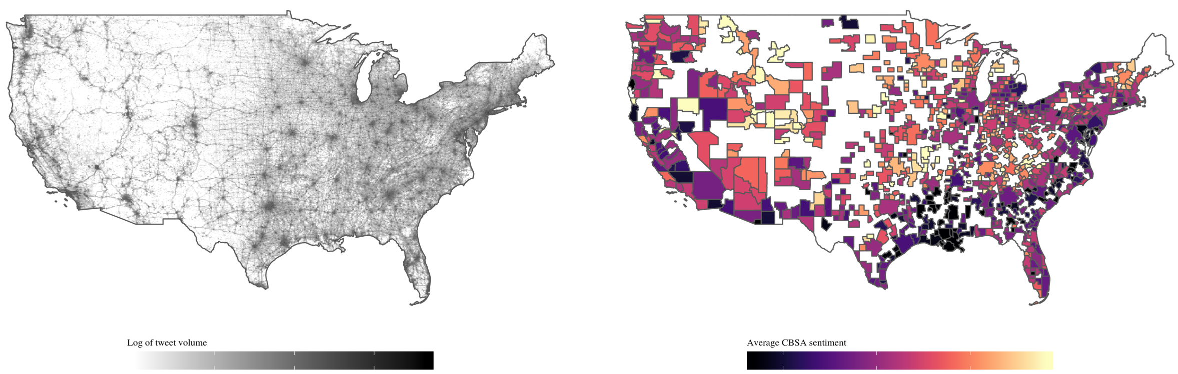

Figure 1: Tweet density and average sentiment by CBSA

Notes: Left panel: pixel shading represents log (base 10) of count of tweets in sample. Right panel: Mean standardized Vader score by CBSA for CBSAs with more than 100 tweets in sample.

The translation of unstructured text data into in a quantitative measures is known as "natural language processing", or NLP. Within NLP, the set of techniques designed specifically to quantify expressed sentiment is known as "sentiment analysis." At the time of this writing, there are more than fifty publicly available algorithms and/or wordlists used to conduct sentiment analysis (Medhat, Hassan, and Korashy 2014). Because the method by which these measures are constructed can differ substantially, analyses using expressed sentiment should ideally demonstrate reasonable consistency across multiple measures. In this paper, I translate tweet content into four measures of expressed sentiment derived from prior work: Hedonometer (Dodds and Danforth 2010), LIWC (Pennebaker et al. 2015), AFINN (Nielsen 2011), and Vader (Gilbert and Hutto 2014). The underlying machinery for each measure is similar: each contains a word list, or dictionary, which contains sentiment scores that correspond to English-language words. The overall measure of sentiment in a piece of text is simply the unweighted average of all scored words within that piece of text.

Panel A in Table 1 describes the unstandardized sentiment measures in the sample, although I standardize the measures prior to analysis for comparability. Following prior work, I pre-process each tweet before scoring in order to increase the precision of the NLP algorithms (Pak and Paroubek 2010). I remove punctuation, URLs, hashtags (e.g., "#job"), and mentions (e.g., "@person") to isolate the word selection in the tweet. Because the independent variable of interest is weather, I remove tweets that contain any weather-related terms (see Table A10 for the list of weather teams I exclude) to ensure that the responses do not capture the sentiment of observations about the weather, only changes in general sentiment due to weather. Once the tweets have been pre-processed, I score them for sentiment using the pre-existing dictionary (AFINN, Hedonometer, and LIWC), with the exception of Vader, which contains its own pre-processing routines. The online appendix gives background and additional detail for each measure. Finally, in addition to the sentiment measures I include a profanity measure intended to capture the use of online vulgarity, changes in which are reported as a percentage of average profanity used in the sample. Figure 2 shows state average expressed sentiment, ordered from lowest to highest. While cross-sectional differences in expressed sentiment provide limited causal insights, the plot suggests that colder regions express higher sentiment on average: the top six states are Montana, New Hampshire, Vermont, Wyoming, Minnesota, and South Dakota.

Table 2 shows the correlations between the five measures at the CBSA-day level. All of the measures are strongly positively correlated with each other, except the measure of profanity which is negatively correlated with all measures.

Table 2: Measure correlations: CBSA-date means

	AFINN-111	Hedonometer	LIWC	Vader	Profanity
AFINN-111	1.00	0.65	0.73	0.78	-0.58
Hedonometer		1.00	0.59	0.73	-0.35
LIWC			1.00	0.77	-0.38
Vader				1.00	-0.40
Profanity					1.00

Notes: Pairwise correlations of CBSA-date means of measures of standardized expressed sentiment and profanity measure.

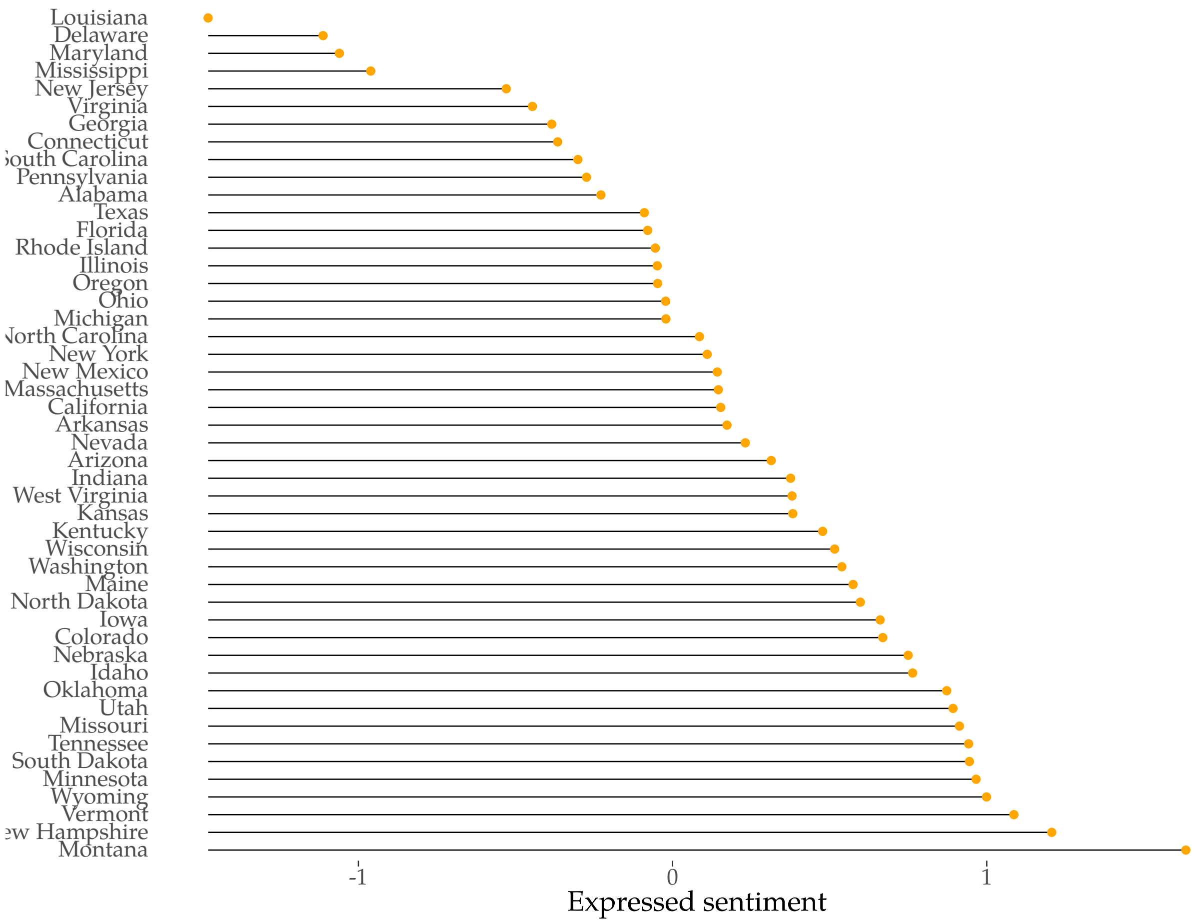

Figure 2: Expressed sentiment by state

Notes: State averages for Vader.

Next, Table 3 documents the correlation between annual state-level averages for each measure and compares them to the annual state-level Gallup index of subjective well-being.[7] In aggregate, the four sentiment measures are highly correlated. The last row compares the measures to state-level measures of well-being estimated by Gallup in 2016. These correlations are positive (except for the profanity measure, as expected) but relatively low (they range between 0.20 and 0.37), indicating that expressed sentiment does not capture the saame underlying object as subjective well-being.

This finding emphasizes the importance of controlling for geographic differences to account non-uniformity in language and tone across space. The measures capture notable geographic heterogeneity: the right panel of Figure 1 documents average sentiment (as measured by Vader) by CBSA across the country. Two patterns are visually evident: urban CBSAs and CBSAs in the northern part of the United States show higher average sentiment

7. Now the "Gallup-Sharecare Well-Being Index", available here: https://wellbeingindex.sharecare.com/.

Table 3: Measure correlations: state-year means

	AFINN-111	Hedonometer	LIWC	Vader	Profanity	Gallup
AFINN-111	1.00	0.88	0.96	0.96	-0.94	0.37
Hedonometer		1.00	0.88	0.89	-0.83	0.20
LIWC			1.00	0.97	-0.89	0.32
Vader				1.00	-0.90	0.31
Profanity					1.00	-0.27
Gallup						1.00

Notes: Pairwise correlations of state annual averages of standardized expressed sentiment and profanity measure and state-level annual Gallup polls of subjective well-being index.

relative to rural CBSAs and CBSAs in the southern part of the country.

2.2 Weather data

To obtain local estimates of daily weather across the contiguous United States, I use the PRISM Climate Group's AN81d gridded weather dataset. These data provide daily measures of minimum temperature, maximum temperature, and precipitation at roughly 4×4 kilometer grid cells for the entire United States. The data are produced using a model that interpolates measurements from more than 10,000 weather stations and corrects for altitude and other influences on local climate (Daly et al. 2002). The second panel in Table 1 describes sample statistics for the PRISM data. I aggregate the gridded data to the CBSA level using population weights to ensure that the weather covariates reflect the average weather experienced by individuals within each CBSA.

Prior work suggests that other weather variables besides temperature and precipitation may be drivers determinants of emotional state (Dennisenn et al. 2008). Accordingly, I also gather daily data on the proportion of day that was overcast, relative humidity, station pressure, and wind speed from 2,162 weather stations included in the NOAA Quality Controlled Local Climatological Data, or QCLCD.

3 Estimating preferences for temperature

3.1 Empirical approach

I identify the causal effect of temperature on expressed sentiment using a panel fixed effects model, with temperature entering the regression using a flexible functional form. This flexibility is justified for the following reasons: first, prior work estimating temperature has documented non-linearities across a wide array of responses to temperature (Carleton and Hsiang 2016), second, an appropriate flexible functional form should reveal the shape of the underlying response function, linear or otherwise (Hsiang 2016) and third, intuition suggests that there is some bliss point for temperature, if only because temperatures which threaten human survival are clearly not preferable. The value of the panel nature of the dataset is that it allows me to control for unobservable cross-sectional and seasonal variation. After accounting for this variation, I follow previous work in the field in interpreting the estimated coefficients as representing the causal effect of temperature on expressed sentiment (Dell, Jones, and Olken 2014).

Specifically, I estimate the following statistical model:

$$\overline{S}_{cd} = f(T_{cd}) + P_{cd} + \phi_c + \boldsymbol{\phi}_{\text{time}} + \varepsilon_{cd} \tag{1}$$

Let c and d index CBSA and date. $\overline{S}_{cd}$ is the CBSA-day average of one of the four measures of sentiment described in Section 2. T_{cd} is the maximum daily temperature in a CBSA, and $f(T_{cd})$ is a flexible function of temperature, which I implement in practice using a binned model specification to allow for nonparametric responses of expressed sentiment to temperature. In particular, I let $f(T_{cd}) = \sum_b^B \beta_b T_{cd}^b$, where T_{cd}^b is an indicator variable equal to one if T_{cd} falls in the given bin b. $P_c d$ is daily precipitation. ϕ_c represents CBSA fixed effects and $\boldsymbol{\phi}_{\text{time}}$ represents a set of additional temporal controls, including month, year, day of week, holiday, and state-specific time trends. ε_{cd} is the idiosyncratic error term, clustered by both CBSA and date. I estimate the model using weighted least squares, where the weights are counts of total scored tweets in a given CBSA.

T_{gd}^b specifies one, three, or five degree bins running between 0 to 40 degrees C, with edge bins for all observations with maximum temperature less than 0 or greater than 40.[8] I include both three and five degree versions of this model as part of the main results I present in the paper, and a comparison of all three bin sizes in the appendix. For all bin widths, I choose the bin that contains 22.5 C as the omitted category. This choice does not alter the shape of the estimated response function, since relative differences between conditional means are preserved, but this choice reflects the prior finding that Americans prefer 65 F (18.3 C) average daily temperature (Albouy et al. 2016). In my sample, because I use daily maximums rather than averages, this corresponds to the omitted bin that I choose.

As shown earlier, the right panel of Figure 1 documents cross-sectional variation in sentiment. Although all regions have a mix of high and low-sentiment CBSAs, visual

8. Because three does not multiply evenly into 40, the upper limit for the three degree bin specification is 39 C.

inspection suggests that there is substantial inter-regional variation in expressed sentiment. Additionally, prior evidence suggests that individuals with higher incomes tend to experience higher levels of life satisfaction and can afford to locate in areas with generally pleasant climate (Easterlin 2001). If this regional variation, which may result from cultural or economic factors, correlates with regional weather differences, a naïve estimate of the relationship between weather and expressed sentiment is likely to be biased. To account for this regional variation in sentiment, I include CBSA fixed effects ϕ_c. These fixed effects ensure that the model is estimated on deviations from CBSA averages rather than on cross-sectional differences in climate, which could correlate with average sentiment or lexical patterns that register as different sentiments. Intuitively, the implication of this modeling choice is that the estimates represent a weighted average of within-CBSA comparisons, e.g., the difference in sentiment in Dane County, WI on a hot day versus a cold day.

A second concern addressed by this identification strategy is the seasonality of both sentiment and temperature. To account for this possibility, I include month of year fixed effects as part of $\boldsymbol{\phi}_{\text{time}}$. Intuitively, this choice of fixed effects implies that the model coefficients represent a weighted average of the differences in sentiment on hot days versus cold days within, e.g., Chicago in June. State time trends and year fixed effects account for potentially correlated trends in both temperature and sentiment that are shared across the sample, while day of week and holiday fixed effects remove statistical noise related to within-week and by-holiday variation in expressed sentiment.

The combination of these fixed effects defines the identification strategy: I assume that deviations in weather are as good as random after accounting for unobserved variation by CBSA, month of year, and year. This assumption is typical of the climate impacts literature (Hsiang 2016), but estimable only in this setting because of the density of the data. The model specified by Equation (1) represents the a defensible tradeoff between minimizing potential bias and maximizing residual variance, although I also estimate alternative specifications with differing sets of fixed effects. Conditional on the assumptions given above, the coefficients of interest β_b can be interpreted as the average change in sentiment resulting from replacing a day in the omitted bin with a day in temperature bin b.

3.2 Findings

Estimates of Equation (1) indicate statistically significant declines in sentiment below 12 C and above 30 C. For expositional clarity, I first present the main result for each sentiment measure in Figure 3. I show that the shape of the response functions is remarkably similar across the different measures of sentiment. Second, Figure 4 estimates a splined model for a single measure, Vader. Third, Table 4 tabulates the response of the Vader measure under a range of different choices of fixed effects.

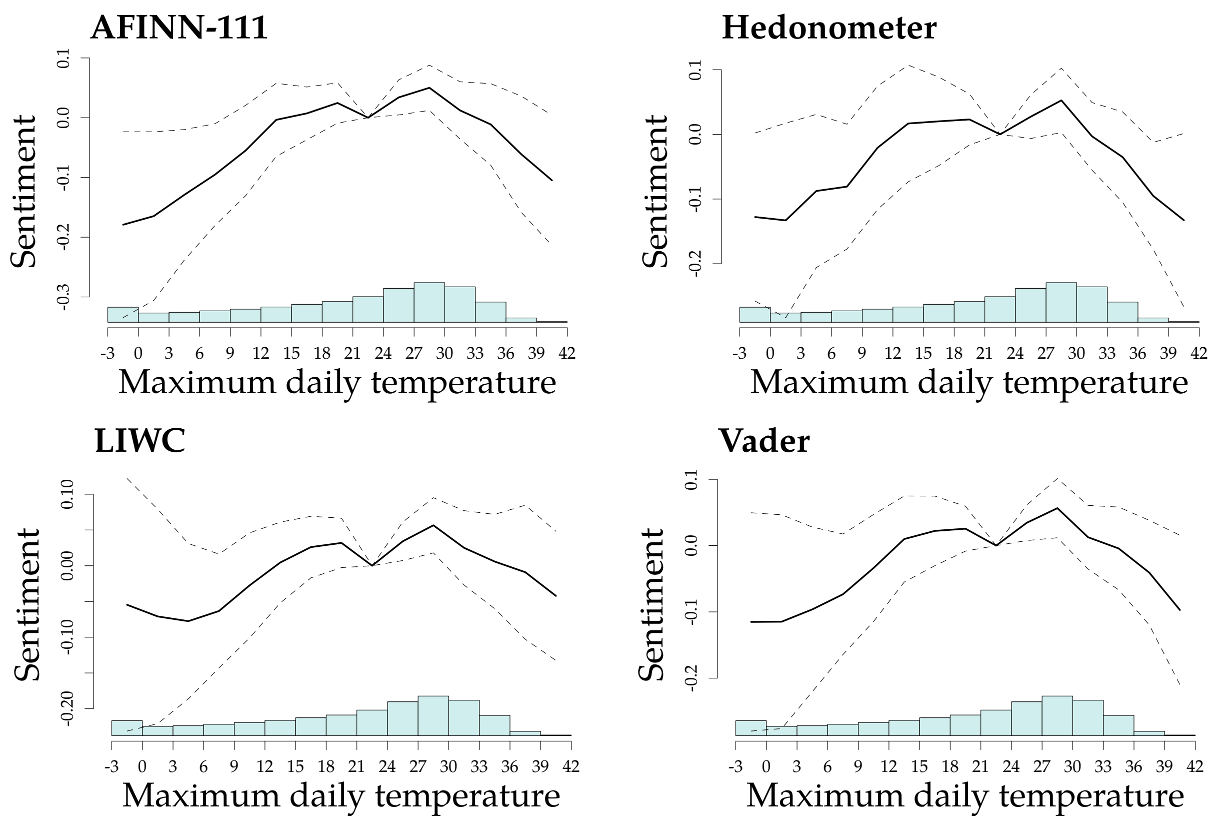

Figure 3: Effect of temperature on expressed sentiment (bins)

Notes: Panels document the temperature response for each of the four standardized measures of sentiment. Solid lines show the regression coefficients on temperature and represent the difference (measured in standard deviations) in CBSA-day sentiment for the temperature bin T_b relative to 21-24 C, controlling for state time trends and fixed effects for CBSA, day of week, holiday, month, and year fixed effects. Dotted lines show 95% confidence intervals estimated using cluster robust standard errors by CBSA and date.

Figure 3 documents the temperature response of all four measures of sentiment estimated using Equation (1). Because each outcome measure is standardized to have mean zero and unit standard deviations, the point estimates β_b represent the change in the conditional mean of expressed sentiment, measured in standard deviations, expected as a result replacing a day with a high of 21-24 C with a day with a high in bin b. I include a histogram underneath each plot to demonstrate the support of the temperature distribution. Each panel includes all four sets of point estimates, with the darker line indicating the central estimate and the dotted lines indicating the 95% confidence interval around that estimate indicating the measure given in the subtitle. The other estimates are included as light gray lines without confidence intervals for comparison.

The upper-left panel documents a decline in the AFINN sentiment measure below 12 C and above 30 C. The difference in sentiment between days with the coldest temperatures (< 3 C) and days in the omitted bin is around 0.15 SD, similar to the difference in sentiment between very hot days (> 39 C) and days in the omitted bin. Confidence intervals are slightly wider for cooler temperature estimates but the point estimates are statistically different from zero at both ends of the temperature range. The AFINN measure estimates the largest cold-weather effect and the second-largest warm weather effect.

The upper-right panel estimates a similar response shape for the Hedonometer measure, both in shape and in magnitude. There is a slight uptick in sentiment at 12-15 C, but this is not a statistically significant difference. Point estimates are statistically different from zero on both ends of the temperature scale. The Hedonometer measure estimates the largest warm-weather effect and the second-largest cool weather effect.

The bottom-left panel documents the response for the LIWC measure. While still within the confidence intervals of the other estimates, LIWC documents the smallest impacts of temperatures on sentiment, with the largest magnitudes slightly less than 0.1 SD. The point estimates have wider confidence intervals and are statistically significant for warm weather temperatures, but less so for cooler temperatures. This difference from the other measures most likely results from the measure's lack of suitability for the microblogging format.

The bottom-right panel documents the response as measured using Vader. Like the AFINN and Hedonometer measures, Vader estimates a statistically significant decline in sentiment below 12 C and above 30 C that reaches about 0.13 SD at maximum. Vader estimates a slightly smaller response than either AFINN or Hedonometer, but a larger response than LIWC.

Each outcome measure in Figure 3 documents a statistically significant negative relationship between sentiment and hot temperatures, relative to a day with moderate temperatures. The magnitudes of the effect sizes differ, ranging from about 0.07 SD to 0.15 SD for the hottest temperature bin. The relationship between sentiment and cold temperatures is slightly less precisely estimated, and one of the four measures fails to reject the null of no difference between cold and moderate temperatures, although the consistent decline of the point estimates provides suggestive evidence of a negative effect in low temperatures. Despite these differences, the results of this exercise are markedly similar across measures: each exhibits the same upside-down "U" shape, each reaches similar magnitudes on both the cold and warm temperature ends of the temperature spectrum, and each is statistically significant at both of those ends (with the exception of LIWC in cooler temperatures).

Because the response functions are consistent across measures, the remainder of the paper focuses on results obtained using Vader. I next replace the binned $f(T)$ in Equation (1) with a flexible spline model. Specifically, I replace $\sum_{b\neq 20\text{-}25}^{B} \beta_b T_{cd}^{b}$ in Equation (1) with a set of basis vectors for a natural spline with knots at the 25th, 50th, and 75th percentile of observed daily maximum temperature in my data. To estimate standard errors, I bootstrap this model 1000 times, which produces the additional benefit of allowing me to estimate the preferred temperature for each run of the model. Figure 4 documents this result; as expected, the shape of the response function in the left panel is similar to that found for the bottom-right panel in Figure 3. The histogram in the right panel documents the preferred temperature for each run of the model, where the median estimate of preferred temperature is at around 22.7 C.

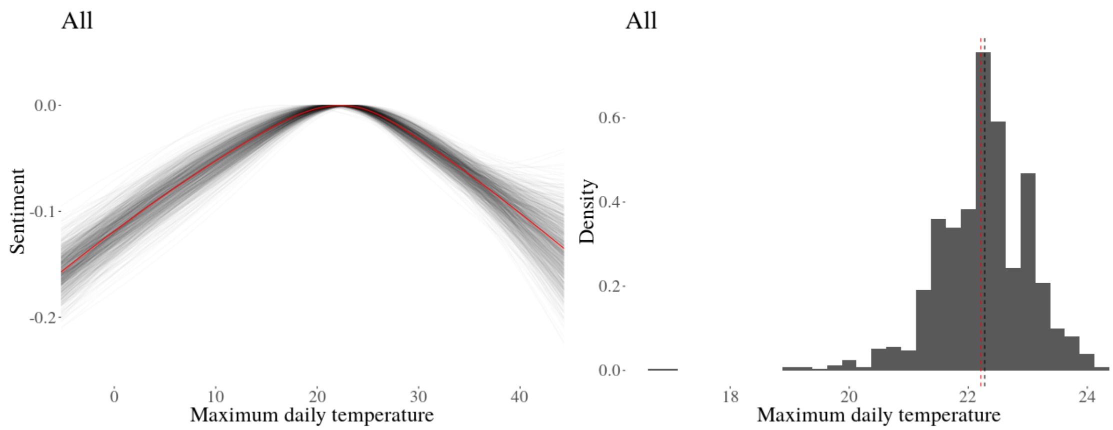

Figure 4: Effect of temperature on expressed sentiment (splines)

Notes: Left panel: documents the response of the expressed sentiment (measured using Vader and in standard deviations) to temperature using a splined model. Darker red line is the splined response function estimated from the entire dataset, with knots at the 25th, 50th, and 75th percentile of experienced daily maximum temperatures. Lighter gray lines are splines estimated similarly using 1,000 bootstrapped samples. All lines are normalized s.t. the highest point of the spline has y = 0. Regressions include CBSA, month, and year fixed effects. 95% confidence intervals clustered by CBSA and date. Right panel: Histogram of estimated preferred maximum daily temperatures for 1,000 bootstrap iterations.

Table 4 estimates the effect of temperature on expressed sentiment using five degree C bins and across a range of choices of fixed effects. All columns include CBSA fixed effects. Column (1) reflects the baseline specification, which includes month and year fixed effects. As in to Figure 3, I observe negative and statistically significant effects below 10 and above 30. Column (2) adds day of week and holiday fixed effects to absorb weekly variation in sentiment (see Figure 9) and variation related to holidays, but the estimated effects are virtually unchanged. Column (4) introduces state-by-month fixed effects to account for regionally distinct seasonal trends. Unlike previous estimates, this does alter the shape of the response function: cooler temperatures are no longer perceived as unpleasant, and the preferred temperature seems to be lower, between 10 and 15 C. Column (4) replaces the state-by-month fixed effects with month of sample fixed effects. Under this specification, the cold temperature estimates are again large and significant. Column (5) replaces the state-by-month fixed effects from column (3) with a more flexible set of state trends and finds largely similar results, although the point estimates are somewhat larger for the cold temperature estimates.

Broadly, I find qualitatively similar results across a range of specifications. Both hot and cold temperatures have a negative effects on expressed sentiment, although only the hot temperature results are statistically significant across the full range of specifications. The appendix also includes additional specification checks, including inclusion of additional weather variables (Table A12), variations on bin width (Figure A11), and user-level estimates (Figure A12) none of which qualitatively alter the baseline results either.

The negative relationship between temperature and sentiment below 12 C and above 30 C resembles that estimated by Albouy et al. (2016), who find that individuals pay to avoid warm temperatures in summer and cold temperatures in winter. The preferred model estimates the magnitude of the difference between a moderate day and an extremely cold or hot day to be about 0.1 SD, or roughly half of the difference in sentiment observed on a Sunday relative to a Monday. A large literature has documented the impact of climate on conflict (Burke, Hsiang, and Miguel 2015a); one possible mechanism is the finding that warm temperatures encourage aggressive behavior (Kenrick and MacFarlane 1986). To understand whether the expressed sentiment response to temperature is due in part to this aggression mechanism, I also estimate the relationship between temperature and expressions of profanity. Using a list of more than 300 profanities, I estimate Equation (1) with the percent of tweets in a CBSA-day that contain a profanity as the outcome of interest. Figure 5 plots the results.

I find that use of profanity rises in both hot and cold temperatures. Previous work on both conflict (Burke, Hsiang, and Miguel 2015a) and on violent crime (Ranson 2014) find that both increase during periods of high temperatures. That I document a similar effect for hot temperatures aligns with the hypothesis that increases temperature induce violence by making individual more aggressive. However, I also find that cold temperatures induce more profane text than moderate temperatures. This finding is in contrast to previous work on temperature and aggressive behavior, which has not typically found an increase in crime or conflict during periods of cooler temperatures (Ranson 2014; Burke, Hsiang,

	(1)	(2)	(3)	(4)	(5)
Maximum daily temperature T					
$T \leq 0$	−0.10	−0.11	−0.02	−0.15	−0.06
	(0.06)	(0.06)	(0.06)	(0.05)	(0.05)
$T \in (0, 5]$	−0.10	−0.11	−0.04	−0.15	−0.06
	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)
$T \in (5, 10]$	−0.07	−0.06	−0.01	−0.09	−0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)
$T \in (10, 15]$	−0.01	−0.01	0.02	−0.02	0.02
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
$T \in (15, 20]$	−0.00	0.00	0.01	−0.01	0.00
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$T \in (25, 30]$	0.01	0.01	−0.01	0.01	0.00
	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
$T \in (30, 35]$	−0.03	−0.03	−0.04	−0.02	−0.05
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
$T \in (35, 40]$	−0.06	−0.06	−0.07	−0.05	−0.08
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
$T > 40$	−0.18	−0.16	−0.12	−0.15	−0.12
	(0.06)	(0.05)	(0.06)	(0.05)	(0.05)
Fixed effects					
CBSA	Yes	Yes	Yes	Yes	Yes
Month	Yes	Yes			Yes
Year	Yes	Yes	Yes		Yes
DOW, Hol		Yes	Yes	Yes	Yes
S×M			Yes		
MOS				Yes	
State trends					Yes

Table 4: Effect of temperature on expressed sentiment (sensitivity)

Notes: Dependent variable is standardized Vader in a CBSA-day. Coefficients represent the change (in standard deviations) of sentiment between a day with maximum temperature in the associated temperature bin and a day with temperature $T \in [20, 25)$, the omitted category. All models include precipitation P, standard errors (in parentheses) clustered by CBSA and date.

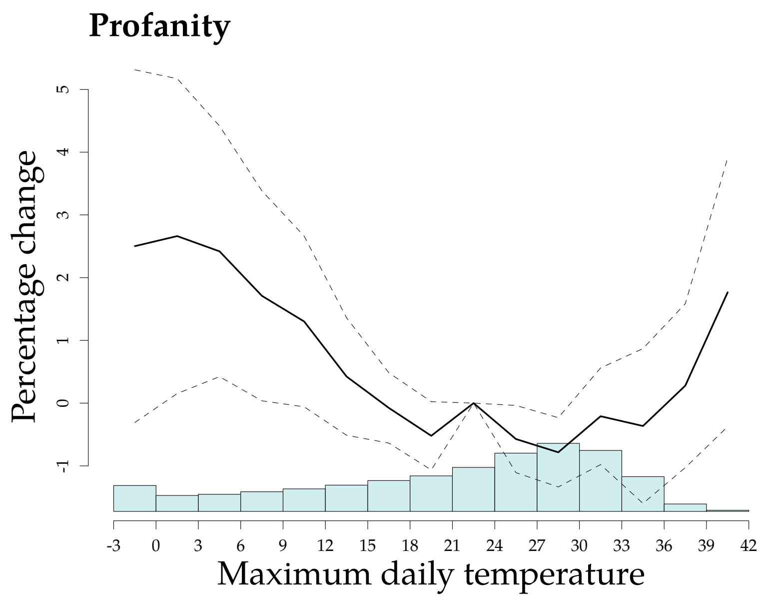

Figure 5: Effect of temperature on profanity

Notes: Figure documents occurrence of profanity in response to temperature, using a list of 300 profanities. Point estimates represent the difference (measured in standard deviations) in CBSA-day profanity occurrence for the temperature bin T_b relative to 21-24 C, conditional on CBSA and state by month of sample fixed effects. 95% confidence intervals estimated using two-way cluster robust standard errors on CBSA and date.

and Miguel 2015a). It may be that aggressive impulses increase in response to temperature discomfort of both kinds, but that cooler temperatures limit opportunities to act on that aggression.

4 Understanding adaptation

As I discuss in Section 1, the climate impacts literature has identified a range of settings in which variation in temperature has had both statistically and economically significant impacts on economic outcomes of interest. The question of whether and to what extent these impacts can be extrapolated to climate change is critically important for projecting cumulative economic impacts. The spread of humans across the planet suggests that, in the long run at least, humans are highly capable of surviving in a wide range of environments. The relatively slow pace of climate change invites the possibility that many of the measured impacts could be partly mitigated by either adaptive responses or by sorting.[9] Empirical estimation of adaptation has presented substantial challenges for researchers working in this area: direct, causally identified models usually rely on long-differences methods as in Burke and Emerick (2015), which in turn rely on sufficient long-run variation in temperature and the outcome of interest across a large geographical area. For most studies, including this one, the requirement of a multi-decadal panel dataset for proper estimation of long-run effects is unattainable. Even for studies with such a dataset available, the research design effectively reduces the number of available observations to the number of observed geographical units, which restricts statistical power and reduces the ability of researchers to strongly reject large portions of parameter space.

As an alternative to providing direct evidence on adaptation or sorting, in this section I take advantage of the richness of the dataset to run a series of empirical tests designed to suggest whether preference adaptation or sorting is likely to occur in this setting. First, I estimate the degree of heterogeneity in temperature-sentiment responses both across the four quartiles average annual temperatures and across the four seasons of the year, finding important differences in the sentiment response across both of these dimensions. Second, I estimate the extent to which these effects persist over time. I find that medium run cold weather effects are consistent with contemporaneous estimates, while the inclusion of longer-run effects of warm temperatures actually negates and then reverses the contemporaneous effect. While these do not, individually or collectively, serve as sharp tests of adaptation, they do help to inform the extent to which temperature preferences do or do not adapt over time.

4.1 Regional heterogeneity in temperature response

Figure 6 estimates separate splined models by region, where each panel identifies the response for the given region. Regions are split by quartiles of average annual temperature, with labels given in order as “Coldest”, “Cold”, “Warm”, and “Warmest”. In order to mitigate the loss of statistical power that results from estimating regional models, I return to the splined model of temperature in Figure 4. To represent standard errors, I bootstrap these estimates with 1000 iterations. For each region, the red line indicates the response identified for the entire sample for that season and the lighter gray lines indicates responses estimated for a set of bootstrapped sample. All lines are normalized such that their maximum value is equal to 0. Clear differences in the sharpness of the sentiment

9. There is a burgeoning literature on understanding adaptation. For more complete discussions of the subject across a range of areas, see Auffhammer et al. (2013), Houser et al. (2014), Graff Zivin, Hsiang, and Neidell (2018), Auffhammer (2013), and Dell, Jones, and Olken (2014).

response to temperature can be observed across regions: colder regions have attenuated responses to cold temperatures, while warmer regions have attenuated responses to warm temperatures. Conversely, colder regions respond more to high temperatures and warmer regions less to cold temperatures. There is one exception: the hottest region has a slightly smaller response to cold temperatures than the second-hottest region, but both responses are more pronounced than that in either the Cold or Coldest regions.

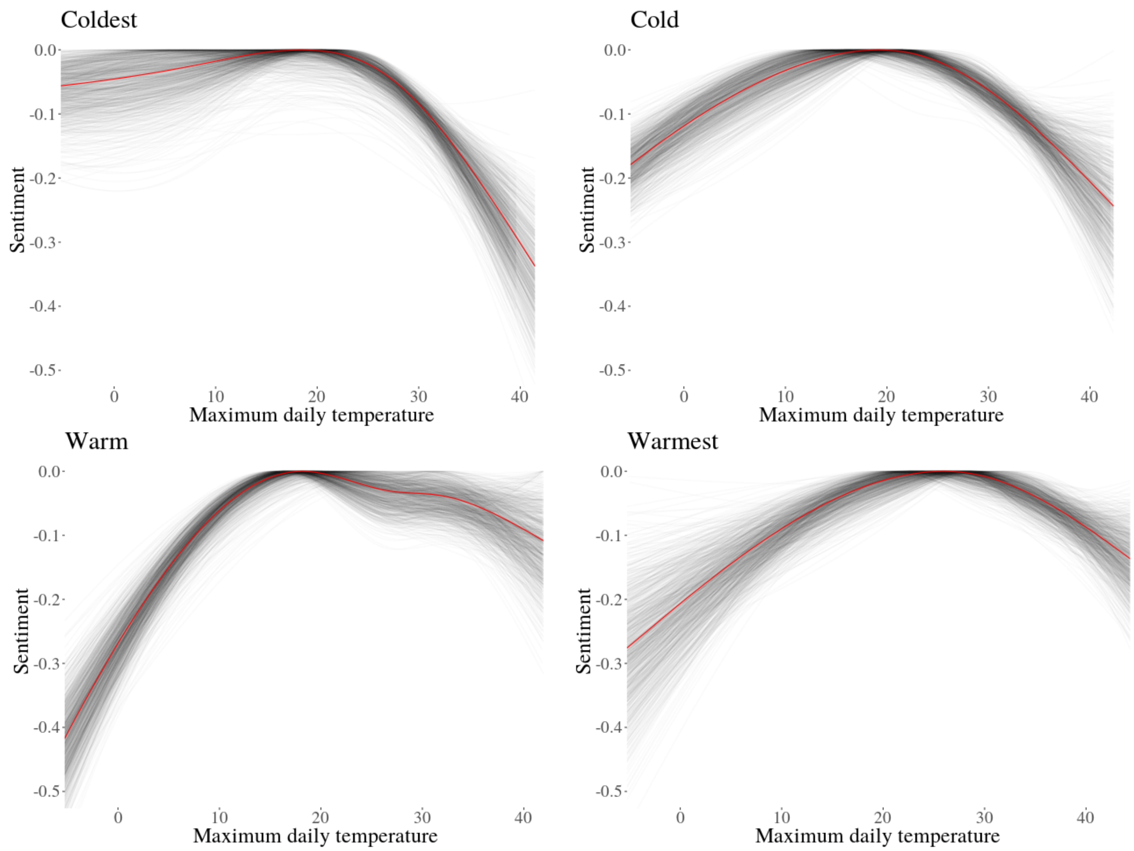

Figure 6: Sentiment responses to temperature differ by region

Notes: Panels document the response of the standardized Vader measure to temperature for each of the four regions, where regions are defined by quartiles of average daily maximum temperature during the sampling frame and are labeled, in order of increasing temperature, "Coldest", "Cold", "Warm", and "Warmest". Dark red line is the splined response function estimated from the entire dataset, with knots at the 25th, 50th, and 75th percentile of experienced daily maximum temperatures. Lighter gray lines are splines estimated similarly using 1,000 bootstrapped samples. All lines are normalized s.t. the highest point of the spline has y = 0. Regressions include CBSA, month, and year fixed effects. 95% confidence intervals estimated using cluster robust standard errors by CBSA and date.

I view this as evidence of either preference adaptation, technological adaptation, or sorting, but I cannot distinguish between these. Individuals may have adapted their preferences to accommodate their climatic zones, they may have a greater degree of technologies available to mitigate extreme temperatures to which they have become accustomed (e.g., air conditioning and indoor heating), or individuals with stronger preferences around lower or higher temperatures may have chosen to vote with their feet, so to speak.

4.2 Seasonal heterogeneity in temperature response

Willingness-to-pay estimates indicate that individuals value warm winters but cool summers (Albouy et al. 2016). This could reflect a stable underlying set of preferences or seasonal shifting of temperature preferences, where the latter would further suggest that adaptive concerns may be in order in this setting. To distinguish between these two possibilities, I estimate the model given in the previous section by season rather than region, where winter is defined as December to February, spring is March to May, summer is June to August, and fall is September to November. Figure 7 documents the response functions by season. For the purposes of interpretation, the height of each line at a given point should be interpreted as the sentiment response to that temperature relative to the preferred temperature for that sample. I find evidence of seasonality in the response: in the summer and fall, individuals have downward sloping preferences for temperatures: cool temperatures are preferred to warm temperatures. In winter, by contrast, the opposite is the case: preferences slope upwards, indicating that warm winter days are preferred. Strength of preferences in spring are more muted: the "bliss point" centers around 21 C, but the declines expressed sentiment in response to either warm or cold temperatures are more modest. Readers should note that the scale of the outcome variable here is larger than that in Figure 3: the strength of these within-season preferences is masked by the aggregation across the year.

The seasonal responses in Figure 7 suggests important seasonal differences in the effect of temperature on expressed sentiment. The negative impact of cold temperatures in Figure 3 seems to be due to the combination of both winter and spring responses, since in neither summer nor fall do I observe a negative response to cooler temperatures. This finding suggests that, if anything, individuals become more sensitive to cold (heat) during typically colder (warmer) seasons, and that adaptation does not seem to occur on a seasonal time frame.

4.3 Cumulative effects of temperature

In this paper I have focused primarily on the contemporaneous, or same-day, effect of temperature on expressed sentiment. Because contemporaneous reactions capture real-time responses of individuals experiencing these various temperatures, they are most

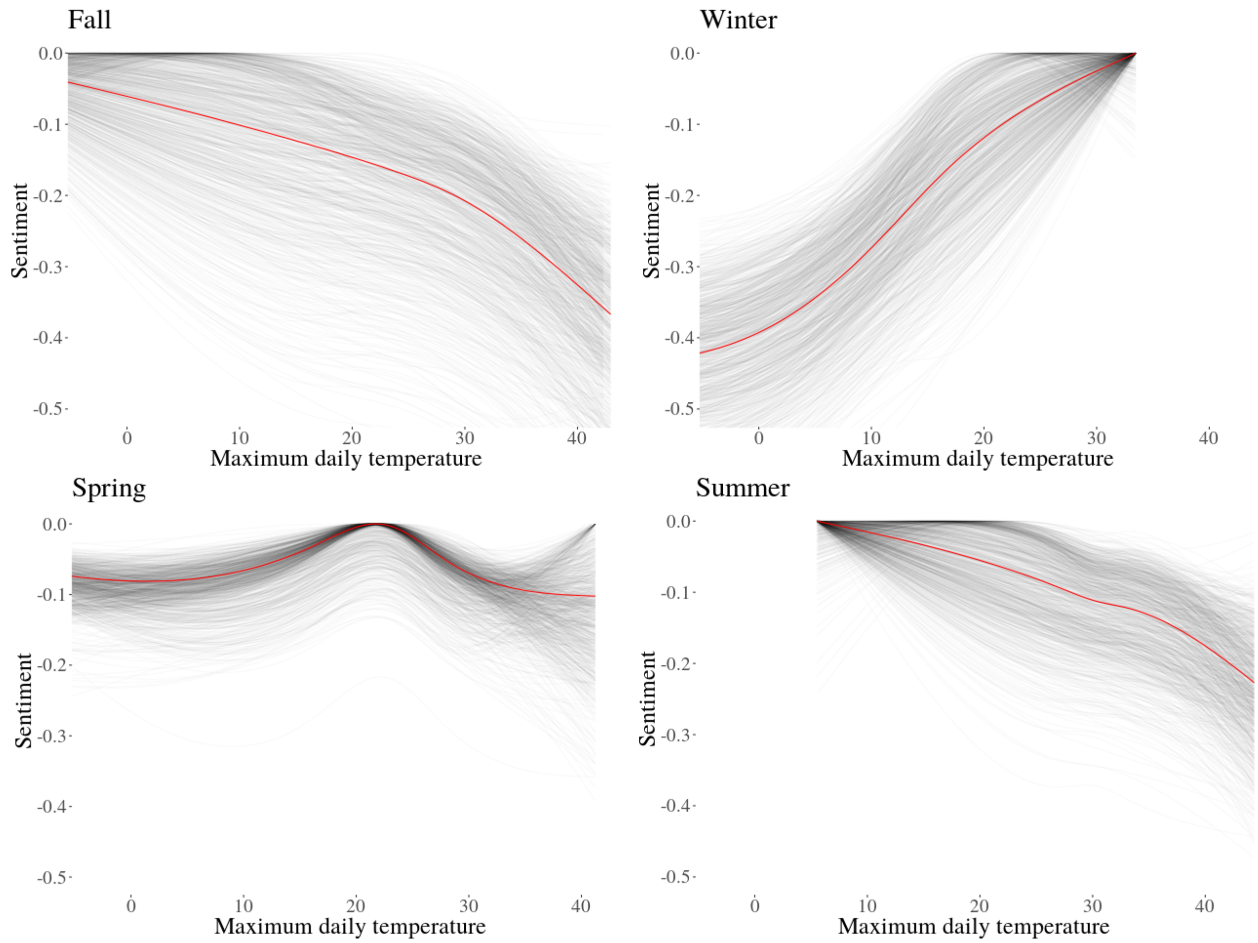

Figure 7: Sentiment responses to temperature differ by season

Notes: Panels document the response of the standardized Vader measure to temperature for each of the four regions, where regions are defined by quartiles of average daily maximum temperature during the sampling frame and are labeled, in order of increasing temperature, “Coldest”, “Cold”, “Warm”, and “Warmest”. Dark red line is the splined response function estimated from the entire dataset, with knots at the 25th, 50th, and 75th percentile of experienced daily maximum temperatures. Lighter gray lines are splines estimated similarly using 1,000 bootstrapped samples. All lines are normalized s.t. the highest point of the spline has y = 0. Regressions include CBSA, month, and year fixed effects. 95% confidence intervals estimated using cluster robust standard errors by CBSA and date.

likely to reflect underlying preferences, such as those that might be derived using an ex ante willingness-to-pay elicitation.

However, it is possible that these effects extend beyond the day in which they are experienced, in which case the contemporaneous estimate could under- or over-estimate the impact of a single day’s temperature. To better understand this impact, I estimate a dynamic cumulative lag model designed to elicit longer-run effects of temperature on expressed sentiment. This model is a set of regressions that begin with Equation (1) and incorporate a progressively increasing number of lags of $f(T)$. The sum of temperature coefficients over the included lags allows me to estimate increasingly longer-run effects of temperature on expressed sentiment. If the sum of the coefficients is stable at the level of the estimated contemporaneous effect (with zero lags in the regression), then the impact of a single day's temperature increase is only felt on that day, and has no subsequent effect. If the sum of the coefficients changes with the inclusion of more lags, then a change in single day's temperature also has an effect on the days following: an increase in magnitude indicates an increasing impact over time, while a decrease in magnitude indicates that subsequent effects may be "washing out" or negating the contemporaneous impact.

Figure 8 plots the results of this dynamic cumulative lag model. While both cold (heating degrees) and warm (cooling degrees) contemporaneous impacts are negative, the inclusion of more lags reveals that while the cold temperature effects are stable, the contemporaneous warm temperature effects are washed out by their lagged impact. In other words, while the impact of a hot day today on expressed sentiment is negative, the impact of a hot day today on expressed sentiment tomorrow is positive, as well as the effect on the day after, and the day after that. [10]

Short-run responses to temperature hold an upside-down U shape, while longer run responses tend to be linearly increasing in temperature until the high 30s. Very hot temperatures remain less preferable.

More broadly, the regional heterogeneity in responses suggests that over long time periods individuals in hotter or colder parts of the U.S. have become partly inured to short-run variations in temperature, although I cannot distinguish whether this has been due to preference adaptation, technological adaptation, or geographic sorting. The seasonal heterogeneity suggests that preference adaptation does not occur seasonally, and that if anything a preference for novelty outweighs adaptation that might be occurring. Finally, the cumulative estimates imply that warm temperatures are contemporaneously dispreferred but have positive effects on sentiment a few days after they are felt. This is not sufficient evidence to demonstrate short-run adaptation per se, but it does suggest that the interactions between contemporaneous temperature and the history of experienced temperature are likely to be important in understanding the amenity value of climate change.

10. Figure A13 shows estimates a monthly, rather than daily, model, and also finds positive impacts of warm temperatures on sentiment, with a similar finding for the hottest bin.

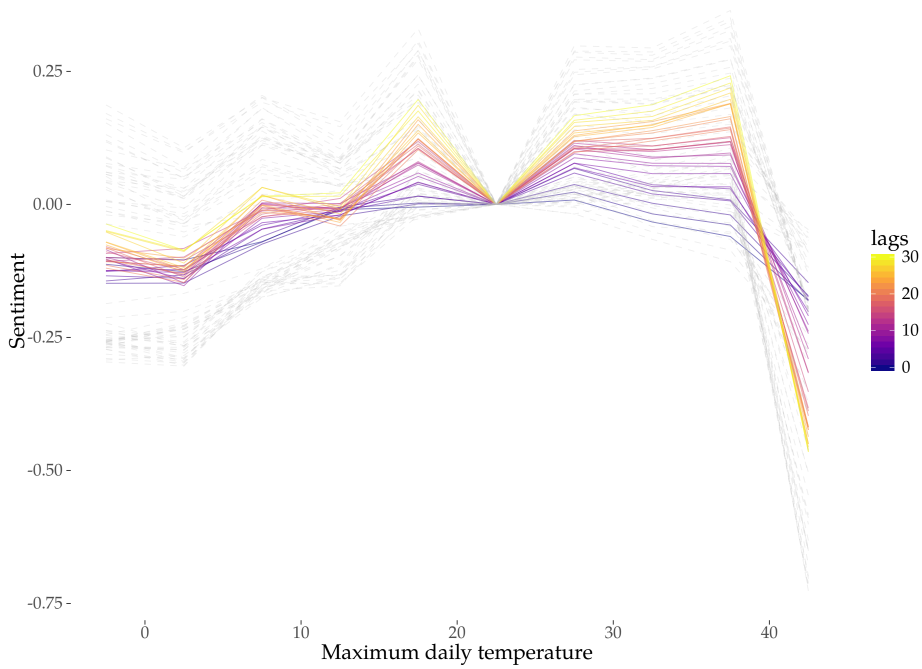

Figure 8: Dynamic changes in temperature impacts over time

Notes: Dynamic cumulative lag model estimated using five degree C bins. Outcome variable is standardized measures of expressed sentiment. Each line represents one set of summed bin coefficient estimates for M lags, where $M \in 0, \ldots, 30$ is the number of lags of binned temperature. Increasingly light coloring indicates more lags included. In addition to contemporaneous and lagged daily maximum temperatures, regressions include precipitation, CBSA, month, and year fixed effects. Standard errors clustered by CBSA and date.

5 Valuing changes in expressed sentiment

By using expressed sentiment as a proxy for underlying preferences for temperature, I am able to mitigate the identification concerns that arise when hedonic or discrete choice models to estimate the value of climate, as described above. The dataset I construct also allows me to estimate the underlying relationship between temperature and expressed sentiment non-parametrically, region- and season-specific response, and cumulative effects. These benefits must be weighed against a significant drawback of this approach: the challenge of interpretation. This is a problem also faced by work using measures of life-satisfaction as the outcome of interest: how much is one unit of expressed sentiment or reported life satisfaction worth?

The advantage of the hedonic and discrete choice approaches is that the derivation of a dollar value for preferences, is straightforward (Albouy et al. 2016; Sinha, Caulkins, and Cropper 2018). By contrast, while I am able to intuitively calibrate the magnitude of the estimates presented here using comparisons to within-week variation in estimated emotional state, backing out estimate of the average willingness-to-pay for changes (or the lack of changes) in temperature is more challenging. However, doing so is important for several reasons: first, assigning a monetary value grounds the size of these effects in a metric that is more likely to be consistently interpreted by different readers; second, monetary calibration of the effect of changes in temperature on emotional state allows researchers and policy analysts to compare the size of these estimates to other documented effects of climate change; third, monetary estimates are critical for inclusion in the three Integrated Assessment Models currently used by the United States Government to estimate the social cost of carbon (Rose 2014).

In the following section I present three separate empirical exercises designed to calibrate the magnitude of the impact of temperature on expressed sentiment that I observe. I demonstrate sentiment impacts for the following four sources of variation: hurricane impacts on affected municipalities, football outcomes for nearby teams, and receipt of parking and speeding tickets.

One possible approach to this challenge is to follow Levinson (2012), who converts reported life satisfaction into a dollar value by dividing the response of life satisfaction to pollution levels with the the response of life satisfaction to cross-sectional differences in income. The drawback of this approach is that it is re-introduces the same cross-sectional concerns in the estimation of the denominator. In my data, examination of the bottom panel of Figure 1 reveals that the low relative expressed sentiment levels of southern and more rural areas would be likely to drive a positive correlation between income and sentiment.[11] However, the case for this as a purely causal relationship is weak, since demographic differences and sorting could easily produce a similar pattern.[12]

The ideal experiment would randomly give money to individuals and monitor the resulting change in sentiment. However, the noise in sentiment expression would require a large sample size and, as a result, a large budget. As an alternative, I construct a class of five distinct validation measures designed to identify impact of a range of different type of societal events as a set of comparisons for the main results. The first validation measure is straightforward: I estimate how sentiment changes over the course of the week. The second validation estimate I use examines the effect of changes in the stock prices on expressed sentiment. Third, I identify the impact of hurricanes on the expressed sentiment of communities impacted by the hurricane. Fourth, I look at the extent to which football outcomes affect their home cities' sentiment. Finally, I subsample my data to examine only users who are given a parking ticket and estimate the sentiment response to this event.

5.1 Validation measure: Day of week

To better understand the measures, I conduct a validation exercise that examines how sentiment changes over the course of the days of the week. First, Figure 9 shows the standardized measures by day of week. The weekly variation in matches prior work (Dodds et al. 2011) and common intuition: weekends and Fridays are preferred to non-Friday weekdays, with the lowest measures of affect occurring on Mondays and the highest on Saturdays. To calibrate to the results shown earlier, note that the average difference in sentiment measure between Sunday and Monday is approximately 0.1σ across measures, or roughly the difference between experiencing a day with maximum temperature between 20 and 25 C and a day with maximum temperature above 40 C.

5.2 Validation measure: Hurricanes

The high winds, heavy rains, storm surges, and the distress and uncertainty hurricanes create results in both substantial economic losses and difficult-to-quantify human hardship (Hsiang and Jina 2014). And although hurricanes mostly affect areas on the Eastern seaboard between June and November, their appearance and path of destruction tend to both be unpredictable in the short term. For these reason, estimating the impact of

11. Cross-sectional estimates of the "effect" of temperature on sentiment confirm this conjecture.

12. There are nearly countless possible candidates for confounding or reverse causality here. A few examples: cultural factors that drive both happiness and economic productivity, weak labor markets that drive down average incomes and decrease happiness due to unemployment, or the well-known relationship between happiness and employability (Mackerron 2012).

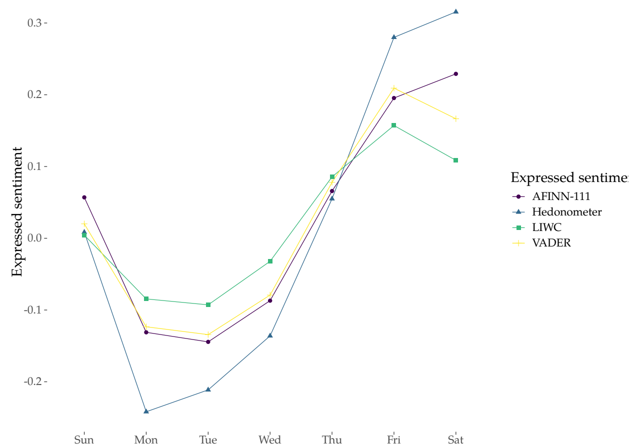

Figure 9: Expressed sentiment by day of week

Notes: Lines represent average standardized measures of expressed sentiment by day of week. Standardization is conducted using the weighted mean and variance of the CBSA-date averages.

hurricanes on expressed sentiment serves as a useful benchmark for the baseline results in this paper. I collect data on hurricane occurrence from the Atlantic Hurricane Database (Landsea and Franklin 2013) and combine it with CBSA-date averages of expressed sentiment, measured using the Vader. To identify the causal impact of a nearby hurricane on expressed sentiment, I estimate the following model:

$$\overline{S}_{cd} = 1[\text{Hurricane in area}]_{cd} + \phi_c + \phi_m + \phi_y + \varepsilon_{cd} \tag{2}$$

Table 5 documents the results of this estimation for all hurricanes (first two columns) and for hurricanes with an average wind speed greater or equal to 40 m/s (second two columns), using two different sets of fixed effects.

On average, CBSAs experience a daily reduction in expressed sentiment of nearly 0.4 standard deviations from any nearby hurricane, while high speed hurricanes cause a reduction of around 0.7 standard deviations. This effect is of the same sign and between 2 and 7 times are large as the effect estimated in Figure 3, suggesting that the impact of nearby hurricanes is more pronounced than the impact of extremely hot or cold temperatures.

5.3 Validation measure: Football game outcomes

I estimate a similar model to Equation (2) using the outcomes of NFL football games during my sample, following in the spirit of Card and Dahl (2011), who shows that unexpected

Table 5: Impact of hurricanes on expressed sentiment

	All	All	40+	40+
Hurricane in area	−0.38	−0.39	−0.70	−0.68
	(0.16)	(0.19)	(0.31)	(0.22)
Fixed effects				
CBSA	Yes	Yes	Yes	Yes
DOW	Yes	Yes	Yes	Yes
Month	Yes	Yes		
Year	Yes	Yes		
MOS			Yes	Yes

Notes: Table documents impact of nearby hurricanes expressed sentiment for nearby CBSAs (metro areas). Outcome variable is standardized CBSA daily average of expressed sentiment. All regressions control for day of hurricane along with listed fixed effects. First and second columns include all hurricanes, third and fourth columns only includes those with a wind speed of 40 m/s. Standard errors clustered by CBSA and date.

football losses correlate with family violence. I map each CBSA to a nearby football team, where being within either 100 or 25 kilometers of a team's home stadium qualifies as "nearby." The model I estimate also includes an indicator for whether the team was playing on a given day. Table 6 documents the findings, where the first two columns take a nearby team as one within 100 km and the second two take a nearby team as one within 25 km, with two different sets of fixed effects for each.

As anticipated, a nearby team suffering a loss causes a negative impact on expressed sentiment. The size of the impact is roughly equal to the negative impact of a very hot day in Figure 3, and is slightly larger when estimated using teams that are within 25 km relative to 100 km.

5.4 Validation measure: Parking and speeding ticket response

I identify more than 8,000 instances where individuals in my sample received parking or speeding tickets. Using only individuals who had at least ten tweets in the seven and after the ticket, I document the sentiment response to receiving a ticket using an event study. Intuitively, this approach compares tweets from users who received a ticket shortly before and shortly after they received in order to control for cross-sectional variation (since users who receive tickets may be different from other users) and time-series trends (since

Table 6: Impact of football outcomes on expressed sentiment

	100 km	100 km	25 km	25 km
Nearby team won	0.10	0.10	0.08	0.08
	(0.03)	(0.03)	(0.05)	(0.05)
Fixed effects				
CBSA	Yes	Yes	Yes	Yes
DOW	Yes	Yes	Yes	Yes
Month	Yes	Yes		
Year	Yes	Yes		
MOS			Yes	Yes

Notes: Table documents impact of football outcomes on expressed sentiment for nearby CBSAs (metro areas). Outcome variable is standardized CBSA daily average of expressed sentiment. All regressions control for day of game along with listed fixed effects. First and second columns include CBSAs within 100 km of a football team's stadium, third and fourth columns include CBSAs within 25 km. Standard errors clustered by CBSA and date.

users who become more like recieve a ticket may be undergoing confounding behavioral changes). Specifically, I estimate:

$$S_{it} = \sum_{k=-7}^{K=7} \beta_k 1[\text{Date}_t - \text{Ticket date}_i + k] + \text{Trend}_t + \phi_s + \varepsilon_{it} \tag{3}$$

where i identifies the user, and t is the date. The β_k are coefficients reflecting the effect of receipt of the ticket on day k, where k number of days after the ticket was received. The top panel of Figure 10 plots β_k over the entire period before and after the ticket, where the omitted category is all tweets not including in the 30 day window. The bottom panel of Figure 10 plots the cumulative effect of a ticket over time, where each point estimate is $\sum_{k=0}^{K} \beta_k$ for $K \in [0, 7]$.

As expected, the receipt of a ticket causes a negative shock in expressed sentiment, which is most sharply experienced on the day of the ticket receipt, but accumulates over time, eventually resulting in a total loss of 1.27 SD of sentiment by day 7. In Section D, I combine this estimate with the median stated value of a ticket to obtain a "value of expressed sentiment." This approach resembles that taken by Levinson (2012), who values pollution using the cross-sectional relationship between stated happiness and differences in income. In this case, the source of variation is an unexpected financial shock, but

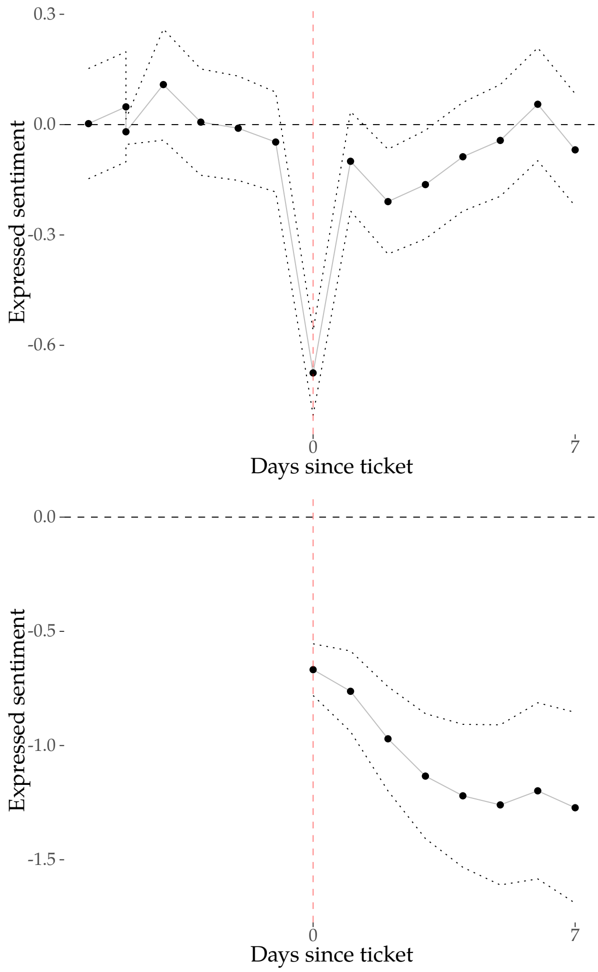

Figure 10: Impact of parking ticket receipt on expressed sentiment

Notes: Left panel: event study estimate of the effect of the receipt of a parking or speeding ticket on standardized Vader sentiment, where receipt of a ticket is self-reported on Twitter. Sample limited to users who received at least one ticket during the sample period, and who had at least 10 tweets in the seven days before and after the ticket receipt. Right panel: estimates from dynamic cumulative lag model. where outcome is standardized Vader sentiment from an increasing number of days since ticket receipt. Point estimates and standard errors are the sum of coefficients on contemporaneous and lagged measures of dummy variable for ticket receipt, with increasing numbers of lags moving from left to right. 95% confidence intervals estimated using cluster robust standard errors by CBSA and date.

because it relies on a highly selected sample (users who experienced parking or speeding tickets during my sample), I present the estimate as a demonstrative exercise in Table A13 and do not here emphasize the willingness-to-pay values it produces.

6 Discussion

By using a contemporaneous responses of expressed sentiment on social media to temperature variation as a proxy for underlying preferences for temperature, I provide an alternative to traditional approaches to non-market valuation. This approach allows me to estimate nonlinear responses of sentiment to temperature and to account for unobserved variation across both space and time. It also allows me to identify spatial and seasonal sources of preference heterogeneity, and to illustrate the difference in short-run and medium-run responses of sentiment to temperature shocks. Finally, the set of validation exercises I demonstrates provides a comparisons for future valuation exercises following this method.

This new approach is not without its drawbacks. The formation of emotional state is highly complex: the physical, biological, and psychological bases for human emotions remain only partly understood (Russell 1980), and the choice of a single dimensional affective scale that is responsive only to changes in levels of utility abstracts away from important nuances regarding the formation of emotion, not to mention its relationship between economic definitions of experienced utility. And although I am able to show that the users who choose to geolocate their tweets are not observationally different than the larger set of Twitter users, extrapoling these results to the general population requires the additional assumption that preferences for climate do not correlate with selection onto the Twitter platform.

Despite these limitations, this paper makes several contributions to the literature. It introduces a new method and data source to estimate preferences for and valuations of public goods while simultaneously accounting for possible unobservable cross-sectional and seasonal variation. It reveals previously unobservable geographic, seasonal, and temporal dynamics of preferences for temperature and provides suggestive evidence of adaptive capacity in this area. And it demonstrates how NLP and specifically sentiment analysis can enable the econometric analysis of large text-based datasets and suggests a psychological channel through which other impacts of climate change may operate. It is worth considering that these results are obtained for the United States, where air conditioner ownership is amongst the highest in the world. Speculatively, the relationship between ambient temperature and sentiment may well be more pronounced in other countries, although cultural differences in temperature sensitivity could mediate this impact. Broadly, this work provides additional evidence that changes in the amenity value of climate is an important component of the overall costs of climate change.

References

Albouy, David, Walter Graf, Ryan Kellogg, and Hendrik Wolff. 2016. "Climate Amenities, Climate Change, and American Quality of Life." *Journal of the Association of Environmental and Resource Economists* 3 (1): 205–246.

Auffhammer, Maximilian. 2013. "Quantifying intensive and extensive margin adaptation responses to climate change: A study of California's residential electricity consumption." *Working Paper.*

Auffhammer, Maximilian, Solomon Hsiang, Wolfram Schlenker, and Adam Sobel. 2013. "Using Weather Data and Climate Model Output in Economic Analyses of Climate Change." *Review of Environmental Economics and Policy* 7 (2): 181–198.

Bentham, Jeremy. 1789. "An Introduction to the Principles of Morals and Legislation." In *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation.* Oxford University Press, January.

Burke, Marshall B., and Kyle Emerick. 2015. "Adaptation to climate change: Evidence from US agriculture." *American Economic Journal: Economic Policy.*

Burke, Marshall, Solomon Hsiang, and Edward Miguel. 2015a. "Climate and Conflict." *Annual Review of Economics* 7:577–617.

———. 2015b. "Global non-linear effect of temperature on economic production." *Nature* 527:235–239.

Card, David, and Gordon B. Dahl. 2011. "Family violence and football: The effect of unexpected emotional cues on violent behavior." *Quarterly Journal of Economics* 126 (1): 103–143.

Carleton, Tamma, and Solomon Hsiang. 2016. "Social and economic impacts of climate." *Science* 353 (6304).

Cline, William R. 1992. *The Economics of Global Warming.* Peterson Institute for International Economics.

Cragg, Michael, and Matthew Kahn. 1997. "New estimates of climate demand: evidence from location choice." *Journal of Urban Economics* 42 (2): 261–284.

Daly, Christopher, Wayne P. Gibson, George H. Taylor, Gregory L. Johnson, and Phillip Pasteris. 2002. "A knowledge-based approach to the statistical mapping of climate." *Climate Research* 22 (2): 99–113.

Dell, Melissa, Benjamin F Jones, and Benjamin A Olken. 2014. "What do we learn from the weather? The new climate-economy literature." *Journal of Economic Literature* 52 (3): 740–98.

Dennisenn, J., Ligaya Butalid, Lars Penke, and Marcel A.G. Van Aken. 2008. "The effects of weather on daily mood: A multilevel approach." *Emotion* 8 (5): 662–667.

Deryugina, Tatyana, and Solomon Hsiang. 2014. "Does the Environment Still Matter? Daily Temperature and Income in the United States." *NBER Working Paper.*

Deschênes, Olivier, and Michael Greenstone. 2011. "Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US." *American Economic Journal: Applied Economics* 3 (4): 152–185.

Diaz, Delavane B. 2014. "Evaluating the Key Drivers of the US Government's Social Cost of Carbon: A Model Diagnostic and Inter-Comparison Study of Climate Impacts in DICE, FUND, and PAGE." *Working Paper.*

Diener, Ed. 2000. "Subjective Well-Being." *American Psychologist* 55 (1): 34–43.

Dodds, Peter Sheridan, and Christopher M. Danforth. 2010. "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents." *Journal of Happiness Studies* 11 (4): 441–456.

Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter." *PloS one* 6 (12): e26752.

Dolan, Paul, Tessa Peasgood, and Mathew White. 2008. "Do we really know what makes us happy? A review of the economic literaure on the factors associated with subjective well-being." *Journal of Economic Psychology* 29 (1): 94–122.

Easterlin, Richard A. 2001. "Income and Happiness: Towards a Unified Theory." *The Economic Journal* 111 (473): 465–484.

Feddersen, John, Robert Metcalfe, and Mark Wooden. 2012. "Subjective Well-Being: Weather Matters; Climate Doesn't." *SSRN Electronic Journal*, no. 627.

Gilbert, CJ, and Eric Hutto. 2014. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In *Eighth International Conference on Weblogs and Social Media.*

Graff Zivin, Joshua, Solomon Hsiang, and Matthew Neidell. 2018. "Temperature and human capital in the short and long run." *Journal of the Association of Environmental and Resource Economists* 5 (1): 77–105.

Hannak, Aniko, Eric Anderson, Lisa Feldman Barrett, Sune Lehmann, Alan Mislove, and Mirek Riedewald. 2012. "Tweetin'in the rain: Exploring societal-scale effects of weather on mood." In *Sixth International AAAI Conference on Weblogs and Social Media.*

Houser, Trevor, Robert Kopp, Solomon M. Hsiang, Michael Delgado, Amir Jina, Kate Larsen, Michael Mastrandrea, Shashank Mohan, Robert Muir-Wood, and D. J. Rasmussen. 2014. *American Climate Prospectus: Economic Risks in the United States.*

Hsiang, Solomon. 2010. "Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America." *Proceedings of the National Academy of Sciences* 107 (35): 15367–15372.

———. 2016. "Climate Econometrics." *Annual Review of Resource Economics.*

Hsiang, Solomon, and Amir Jina. 2014. "The Causal Effect of Environmental Catastrophe on Long-Run Economic Growth." *NBER Working Paper.*

Interagency Working Group on Social Cost of Carbon. 2013. *Technical Update of the Social Cost of Carbon for Regulatory Impact Analysis Under Executive Order 12866.* Technical report. United States Government.

IPCC. 2014. *IPCC Fifth Assessment Report.* Technical report. Cambridge, United Kingdom and New York, NY, USA: IPCC.

Kahneman, Daniel, and Alan B. Krueger. 2006. "Developments in the Measurement of Subjective Well-Being." *Journal of Economic Perspectives* 20 (1): 3–24.

Kahneman, Daniel, and Robert Sugden. 2005. "Experienced utility as a standard of policy evaluation." *Environmental and Resource Economics* 32 (1): 161–181.

Kenrick, D. T., and S. W. MacFarlane. 1986. "Ambient Temperature and Horn Honking: A Field Study of the Heat/Aggression Relationship." *Environment and Behavior* 18 (2): 179–191.

Landsea, Christopher W, and James L Franklin. 2013. "Atlantic Hurricane Database Uncertainty and Presentation of a New Database Format." *Monthly Weather Review* 141, no. 10 (October): 3576–3592.

Levinson, Arik. 2012. "Valuing public goods using happiness data: The case of air quality." *Journal of Public Economics* 96, no. 910 (October): 869–880.

Lobell, David B, Marshall B Burke, Claudia Tebaldi, Michael D Mastrandrea, Walter P Falcon, and Rosamond L Naylor. 2008. "Prioritizing climate change adaptation needs for food security in 2030." *Science* 319 (5863): 607–610.

Mackerron, George. 2012. "Happiness Economics from 35000 Feet." *Journal of Economic Surveys* 26 (4): 705–735.

Marshall, Alfred. 1890. *Principles of Economics.* 1–323.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal* 5, no. 4 (December): 1093–1113.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose." In *Seventh international AAAI conference on weblogs and social media.*

Nielsen, Finn Årup. 2011. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs." *arXiv preprint arXiv:1103.2903.*

Nordhaus, William D. 1991. "To Slow or Not To Slow: The Economics of the Greenhouse Effect." *The Economic Journal:* 920–937.

Nordhaus, William D, and J.G. Boyer. 2000. *Warming the World: Economic Models of Global Warming.* 232. MIT Press.

Pak, Alexander, and Patrick Paroubek. 2010. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In *LREC,* 10:1320–1326.

Pearce, David. 2002. "An Intellectual History of Environmental Economics." *Annual Review of Energy and the Environment* 27 (1): 57–81.

Pennebaker, James W, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical report.

Ranson, Matthew. 2014. "Crime, weather, and climate change." *Journal of Environmental Economics and Management* 67 (3): 274–302.

Rehdanz, Katrin, and David Maddison. 2005. "Climate and happiness." *Ecological Economics* 52 (1): 111–125.

Rose, S. 2014. *The Social Cost of Carbon: A Technical Assessment.* Technical report. Electric Power Research Institute.

Russell, James A. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology.*

Schlenker, Wolfram, and Michael J. Roberts. 2009. "Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change." *Proceedings of the National Academy of Sciences* 106 (37): 15594–15598.

Sinha, Paramita, Martha L Caulkins, and Maureen L Cropper. 2018. "Household location decisions and the value of climate amenities." *Journal of Environmental Economics and Management* 92:608–637.

Stern, Nicholas. 2006. *The Economics of Climate Change,* 662.

United States Government Accountability Office. 2014. *Development of Social Cost of Carbon Estimates.* Technical report July. United States Government Accountability Office.

Welsch, Heinz, and Jan Kühling. 2009. "Using happiness data for environmental valuation: Issues and applications." *Journal of Economic Surveys* 23, no. 2 (April): 385–406.

ONLINE APPENDIX

A Measures of sentiment

The AFINN measure is constructed using an expert-created dictionary that maps words to measures of emotional state. The AFINN-111 dictionary contains 2,477 words scored using integers between -5 and 5, where -5 indicates negative emotional state and 5 indicates positive emotional state. The dictionary focuses on words that are indicative of emotional state, and was created by Nielsen (2011) to analyze language typically used in microblogging. The dictionary is refined from an earlier dictionary built by psychologists to assess the sentiment of written texts (Bradley and Lang 1999).

The Hedonometer measure is constructed in a similar manner to the AFINN measure, but instead uses a dictionary constructed by Dodds and Danforth (2010). The authors crowd-source a dictionary of more than 10,000 words using Amazon's Mechanical Turk service, which outsources tasks to users who are paid for their time. Users were asked to rate each word on a scale from 1 to 9, where 1 indicated negative emotional state and 9 indicated positive emotional state, and measures were averaged across users to get a single measure for each word. Unlike the AFINN measure, the Hedonometer measure scores most commonly-used words regardless of whether they are likely to be indicative of underlying emotional state.

The LIWC measure uses the Linguistic Inquiry and Word Count (LIWC) dictionary created by Pennebaker et al. (2015). Like AFINN and Hedonometer, LIWC uses a dictionary-based method to score text. LIWC contains a variety of dictionaries developed using human categorizations of words: I focus on the lists of words that indicate positive and negative emotion, respectively. The strength of LIWC is that the word lists relating to positive and negative emotion have been independently validated by outside researchers. For example, Kahn et al. (2007) conduct a set of experiments that test whether individuals' stated emotional states correspond to the emotional state estimated from their writing samples using LIWC, and find that LIWC is a valid measure of measuring emotional state.

The VADER measure is a "a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains" (Gilbert and Hutto 2014). VADER is licensed as open-source and is a normalized, weighted composite score. The lexicon used by VADER is constructed by aggregating ratings from 10 independent human raters. The list of candidate words for the lexicon is constructed from previously existing measures of sentiment and augmented using lexical features frequently observed in online contexts such as emoticons (e.g, ":)"), acronyms (e.g., "LOL"), and slang (e.g., "nah"). The VADER measure also includes a mechanism that incorporates information about the word order and intensifiers included in the sentence, so that "very good" is measured as having a higher valence than "good". The measure has been validated against a variety of ground-truth data and found to outperform other measures (Gilbert and Hutto 2014).

Tables A7 to A10 include examples of the word lists used to construct the AFINN, Hedonometer, and LIWC sentiment scores.

B Comparing geolocated tweets to all tweets

In order to identify the weather conditions during which tweets are created it is necessarily the case that all of the tweets in the sample are geolocated. Since geolocation is enabled by choice, it may be the case that geolocating users are different from non-geolocating users. Because Twitter does not provide demographic data on their users, identifying these observational differences is not straightforward. Instead, I compare the word choices across from both geolocated and non-geolocated tweets as a proxy for demographic comparison. To do so, I separately download a sampling of all tweets from the Twitter Streaming API. Using only English Tweets and removing "Retweets" (repetitions of others' tweets), I document how users who geolocate their tweets differ in their language from the non-geolocating population. Table A11 shows that language use does not vary substantially between these two samples: the top 50 words across each group are very similar, and Kendall's tau rank correlation coefficient for the top 1000 words is 0.74.

The close correspondence between the two rankings suggests that the population of users who choose to geolocate their tweets are not markedly different from the Twitter population, at least in terms of their word choice while on the platform. Along with Dennisenn et al. (2008), who finds that the temperature sensitivity of mood does not respond to observable traits, this indicates that the observed preferences in the geolocated Twitter population for temperature are likely to map reasonably well to the Twitter population at large.

C Empirical checks

In this section I document a series of checks intended to test the robustness of the result to different sample selection criteria and model specifications.

Table A7: AFINN word-score examples

Positive Affect		Neutral Affect		Negative Affect	
superb	5	combat	-1	betraying	-3
thrilled	5	apologizes	-1	agonises	-3
hurrah	5	exposing	-1	destroying	-3
outstanding	5	oxymoron	-1	swindle	-3
breathtaking	5	provoked	-1	abhors	-3
roflcopter	4	limited	-1	humiliation	-3
wowow	4	escape	-1	chastises	-3
rejoicing	4	unconfirmed	-1	victimizing	-3
lifesaver	4	passively	-1	bribe	-3
winner	4	blocks	-1	lunatic	-3
miracle	4	poverty	-1	scandal	-3
triumph	4	attacked	-1	outrage	-3
fabulous	4	gun	1	betrayed	-3
roflmao	4	feeling	1	terror	-3
euphoric	4	intrigues	1	abuse	-3
heavenly	4	alive	1	greenwash	-3
fantastic	4	protected	1	falsified	-3
ecstatic	4	unified	1	douche	-3
funnier	4	relieves	1	agonized	-3
winning	4	fit	1	criminals	-3
masterpiece	4	restore	1	defects	-3
masterpieces	4	relieve	1	idiotic	-3
stunning	4	greeting	1	woeful	-3
godsend	4	yeah	1	acrimonious	-3
lmfao	4	cool	1	nuts	-3
lmao	4	vested	1	swindles	-3
rotflmfao	4	clearly	1	lost	-3

Notes: Raw scores shown. Standardized scores used in analysis. Full list includes 2,477 total word-score mappings and can be obtained here: http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

C.1 Additional weather covariates

Because different aspects of weather are frequently correlated, models that omit a key meteorological driver of a given outcome may induce a bias in the estimates of the included weather covariates (Auffhammer et al. 2013). Because the weather dataset I use includes precipitation as well, I include both temperature and precipitation in Equation (1) in order to avoid absorbing the effect of precipitation on expressed sentiment in the temperature estimates. However, since prior findings indicate that a variety of weather variables can impact stated mood (Dennisenn et al. 2008), I estimate a model with additional

Table A8: Hedonometer word-score examples

Positive Affect		Neutral Affect		Negative Affect	
laughter	8.5	fui	5.08	suicide	1.3
happiness	8.44	gilbert	5.08	terrorist	1.3
love	8.42	hart	5.08	rape	1.44
happy	8.3	hij	5.08	murder	1.48
laughed	8.26	hun	5.08	terrorism	1.48
laugh	8.22	indonesia	5.08	cancer	1.54
laughing	8.2	jo	5.08	death	1.54
excellent	8.18	john	5.08	died	1.56
laughs	8.18	juan	5.08	kill	1.56
joy	8.16	knee	5.08	killed	1.56
successful	8.16	laws	5.08	torture	1.58
win	8.12	listed	5.08	arrested	1.64
rainbow	8.1	manhasset	5.08	deaths	1.64
smile	8.1	marion	5.08	raped	1.64
won	8.1	martinez	5.08	killing	1.7
pleasure	8.08	medicaid	5.08	die	1.74
smiled	8.08	medicine	5.08	jail	1.76
rainbows	8.06	meyer	5.08	terror	1.76
winning	8.04	might	5.08	kills	1.78
celebration	8.02	morgen	5.08	fatal	1.8
enjoyed	8.02	morris	5.08	killings	1.8
healthy	8.02	nas	5.08	murdered	1.8
music	8.02	necessarily	5.08	war	1.8

Notes: Raw scores shown. Standardized scores used in analysis. Full list includes 10,223 total word-score mappings and can be obtained here: `http://hedonometer.org/words.html`.

Table A9: LIWC word examples

Positive emotion	Negative emotion
love	hurt
nice	ugly
sweet	nasty

Notes: LIWC is a commercial product, selected examples are described in Tausczik and Pennebaker (2010). Full list includes 905 total words.

weather covariates compiled from the QCLCD weather station data described in Section 2. To minimize measurement error, I include only CBSAs with a QCLCD weather station

Table A10: Weather-related stopwords

blizzard	frostbite	precipitation
breeze	frosty	rain
chilly	gail	rainbow
clear	gust	showers
clouds	hail	sleet
cloudy	heat	snowflakes
cold	hot	soggy
damp	humid	sprinkle
dew	hurricane	sunny
downpour	icy	thunder
drizzle	lightning	thunderstorm
drought	misty	typhoon
dry	moist	weather
flurry	monsoon	wet
fog	muddy	wind
freezing	overcast	windstorm
frigid	pouring	windy

Notes: Author construction.

present.

Table A12 tabulates the regression results from a model that adds relative humidity, wind speed, air pressure, and the percent of the day that was reported as overcast to Equation (1). The results are qualitatively similar, but document a more dramatic decline in mood in higher temperatures. Relative humidity and % overcast both negatively affect expressed sentiment, but their effects are small relative to the reported change in sentiment resulting from temperature.

C.2 Bin widths

Because statistical models employing bin specifications can sometimes be affected by the selection of bin width, I estimate models with 1, 3, and 5 C bin widths in Figure A11.

C.3 Individual tweet-level analysis

Because Twitter users choose when – and when not – to tweet, the selection mechanism into the sample could induce a compositional bias in the estimates observed in Figure 3, a sample selection effect akin to that described by (Heckman 1979). This can also be viewed as a form of the ecological fallacy: the observation that the properties of aggregated groups may not reflect properties of the individuals in the underlying populations (Robinson 1950). To fix ideas, imagine two types of Twitter users: positive and negative. Positive users create only positively-scored tweets, while negative users create only negative tweets. Because neither type will change the content of its tweets in response to temperature, the

Table A11: Comparison of top word usage across geolocation choice

Term	Rank (non-geolocated)	Rank (geolocated)
im	1	1
just	2	2
like	3	3
one	4	6
dont	5	4
love	6	7
get	7	5
can	8	9
now	9	11
new	10	15
good	11	8
people	12	14
know	13	10
time	14	12
u	15	31
go	16	13
see	17	19
amp	18	17
really	19	22
want	20	30
year	21	16
day	22	20
cant	23	21
need	24	24
got	25	18
video	26	228
think	27	27
thank	28	29
happy	29	23
thats	30	28

Notes: Rank of terms used in non-geolocated versus geolocated tweets. Corpus includes only English tweets that were not retweets delivered through the Streaming API from 2018-11-30 to 2019-12-14.

	(1)	(2)	(3)	(4)
Maximum daily temperature T				
$T \leq 0$	−0.08	−0.06	0.03	−0.09
	(0.06)	(0.05)	(0.06)	(0.05)
$T \in (0, 5]$	−0.08	−0.06	0.02	−0.09
	(0.05)	(0.04)	(0.05)	(0.04)
$T \in (5, 10]$	−0.04	−0.01	0.04	−0.04
	(0.04)	(0.03)	(0.04)	(0.03)
$T \in (10, 15]$	0.00	0.02	0.06	0.01
	(0.03)	(0.03)	(0.03)	(0.03)
$T \in (15, 20]$	−0.01	0.01	0.03	0.00
	(0.02)	(0.01)	(0.01)	(0.01)
$T \in (25, 30]$	−0.00	−0.00	−0.03	−0.00
	(0.02)	(0.01)	(0.01)	(0.01)
$T \in (30, 35]$	−0.06	−0.06	−0.10	−0.06
	(0.02)	(0.02)	(0.02)	(0.02)
$T \in (35, 40]$	−0.09	−0.10	−0.13	−0.08
	(0.03)	(0.03)	(0.03)	(0.03)
$T > 40$	−0.14	−0.14	−0.13	−0.12
	(0.05)	(0.05)	(0.06)	(0.05)
Precipitation	−0.00	−0.00	−0.00	−0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Relative Humidity	−0.00	−0.00	−0.00	−0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Wind Speed	−0.00	−0.00	0.00	−0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Station Pressure	0.05	−0.02	−0.02	0.00
	(0.07)	(0.05)	(0.05)	(0.05)
Overcast	−0.09	−0.10	−0.08	−0.11
	(0.02)	(0.02)	(0.02)	(0.02)
Fixed effects				
CBSA	Yes	Yes	Yes	Yes
Month	Yes	Yes		Yes
Year	Yes	Yes	Yes	Yes
DOW, Hol		Yes	Yes	Yes
S×M			Yes	
MOS				Yes
State trends	Yes	Yes	Yes	Yes

Table A12: Additional weather variables

Notes: Dependent variable is sentiment in a CBSA-day. Coefficients represent the change in standard deviations of sentiment between a day with maximum temperature in the associated temperature bin and a day with temperature $T \in [20, 25)$, the omitted category. Units of air pressure are inches in hundreds, overcast is a variable from zero to one capturing proportion of daytime with overcast sky. Standard errors clustered by CBSA and date.

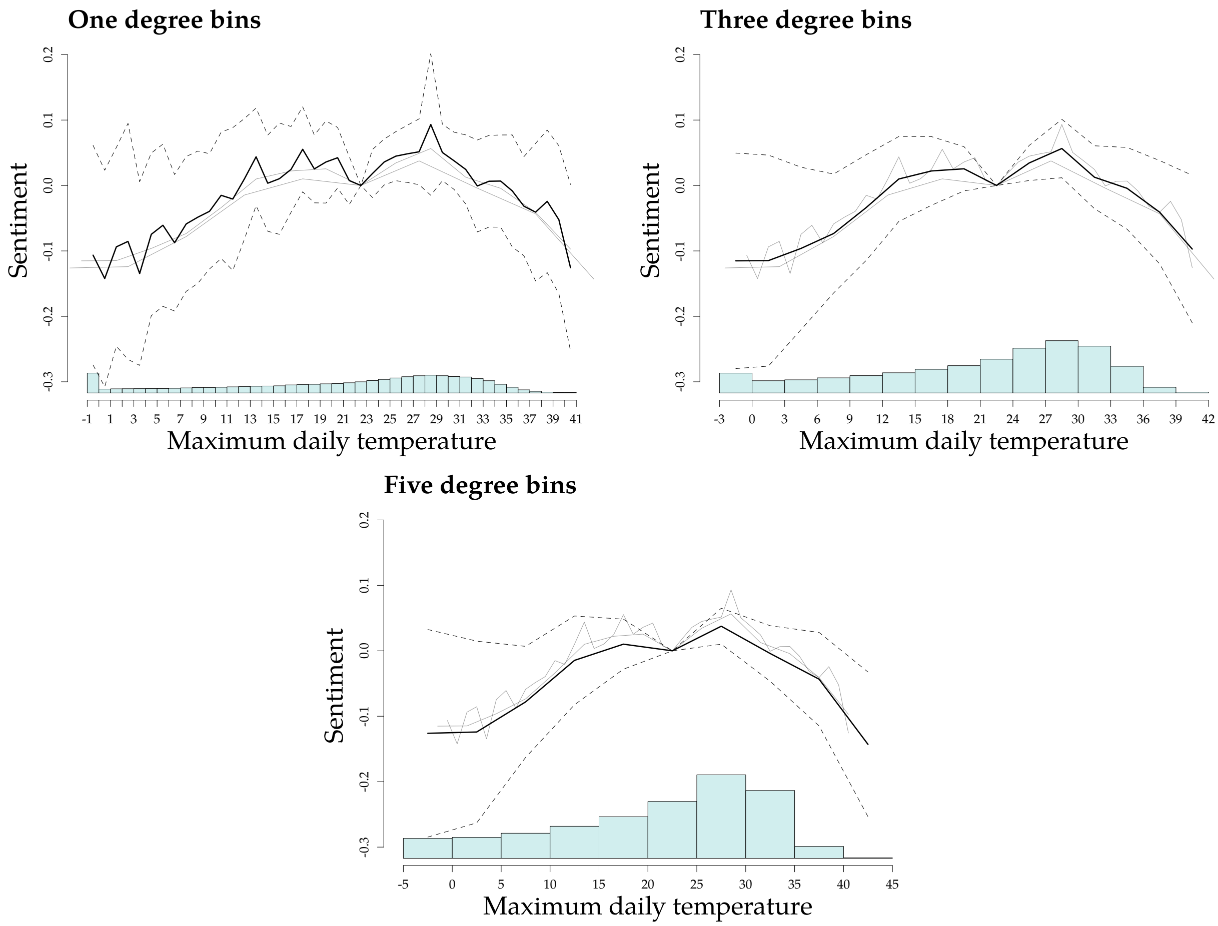

Figure A11: Bin width

Notes: Comparison of the sentiment response to temperature across bin widths of 1, 3, and 5C. Dark solid and dashed lines indicate model estimates and standard errors for the given bin width, gray lines illustrate estimates for other bin widths for comparison. All models include precipitation P, CBSA, month, and year fixed effects. Standard errors clustered by CBSA and date.

true underlying effect of temperature on their sentiment in zero. However, suppose as well that positive users choose to put their phones away when it's very cold or very hot, whereas negative users are unaffected. An econometric approach using CBSA averages that does not control for the type of user will in fact pick up this change in the sampling frame rather than the true effect.

Since the data I collect include an identifier for the tweet creator, I can account for compositional sorting in my sample using user fixed effects. To do so, I estimate the following model:

$$E_{id} = \sum_{b \neq 20\text{-}25}^{B} \beta_b T_{cd}^{b} + \phi_i + \phi_m + \phi y + \varepsilon_{id} \quad (4)$$

This model replaces CBSA fixed effects with user fixed effects ϕ_i in equation Equation (1).

The model requires the use of the disaggregated sample of tweets in my dataset; for computational reasons, I use a 20% subsample of the users with more than 100 tweets in my sample to estimate this model.

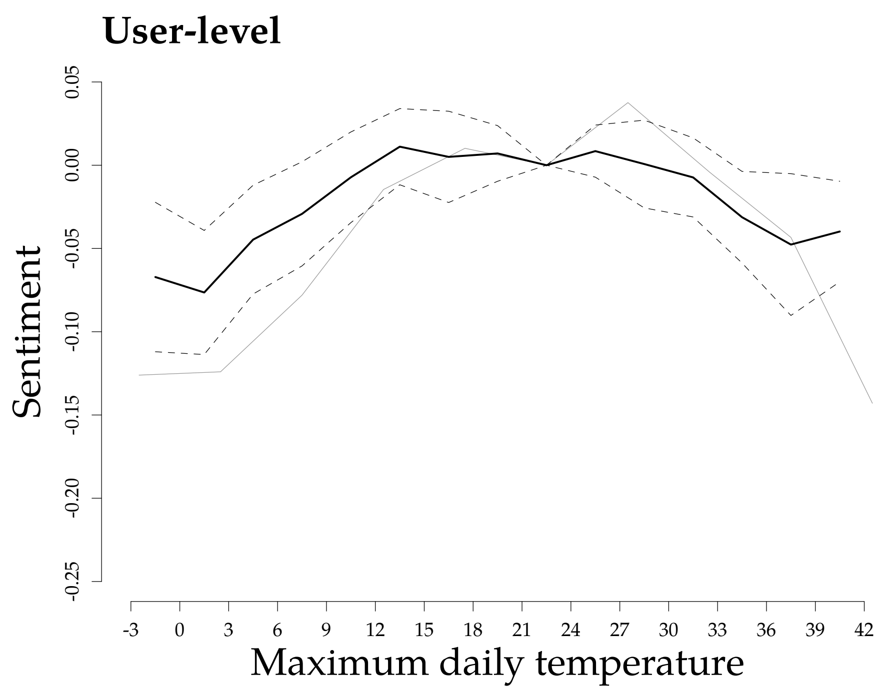

Figure A12: CBSA-level and user-level model comparison

Notes: Plot compares the hedonic response to temperature for models using CBSA and user-level estimates. Dark line indicates response and confidence interval of estimates from binned, user-level model, which replaces CBSA fixed effects in Equation (1) with user fixed effects and is estimate on tweets from a 20% sample of tweets from users with more than 100 overall tweets in sample. The light gray line is the baseline specification using CBSA-day averages as observations, plotted in top left panel of Figure 3. 95% confidence intervals are given for the user-level estimates as the dotted lines, estimated using standard errors clustered by CBSA and date.

To compare the results between the user fixed model and the baseline model, I overlay the estimates from each model in Figure A12. I find qualitatively similar results for the measures, although the estimates for higher temperatures are attenuated in the individual fixed effects model relative to the baseline model. It is possible that this is evidence of some compositional sorting at higher temperatures, but more likely the result of measurement error driven by using a sparser source of variation. The negative response to cold temperature is nearly identical between models, suggesting that the source of the differential is heterogeneous in temperature.

C.4 Monthly model

Figure A13 estimates the main model in the using monthly aggregate date to identify medium-run impacts of temperature on sentiment.

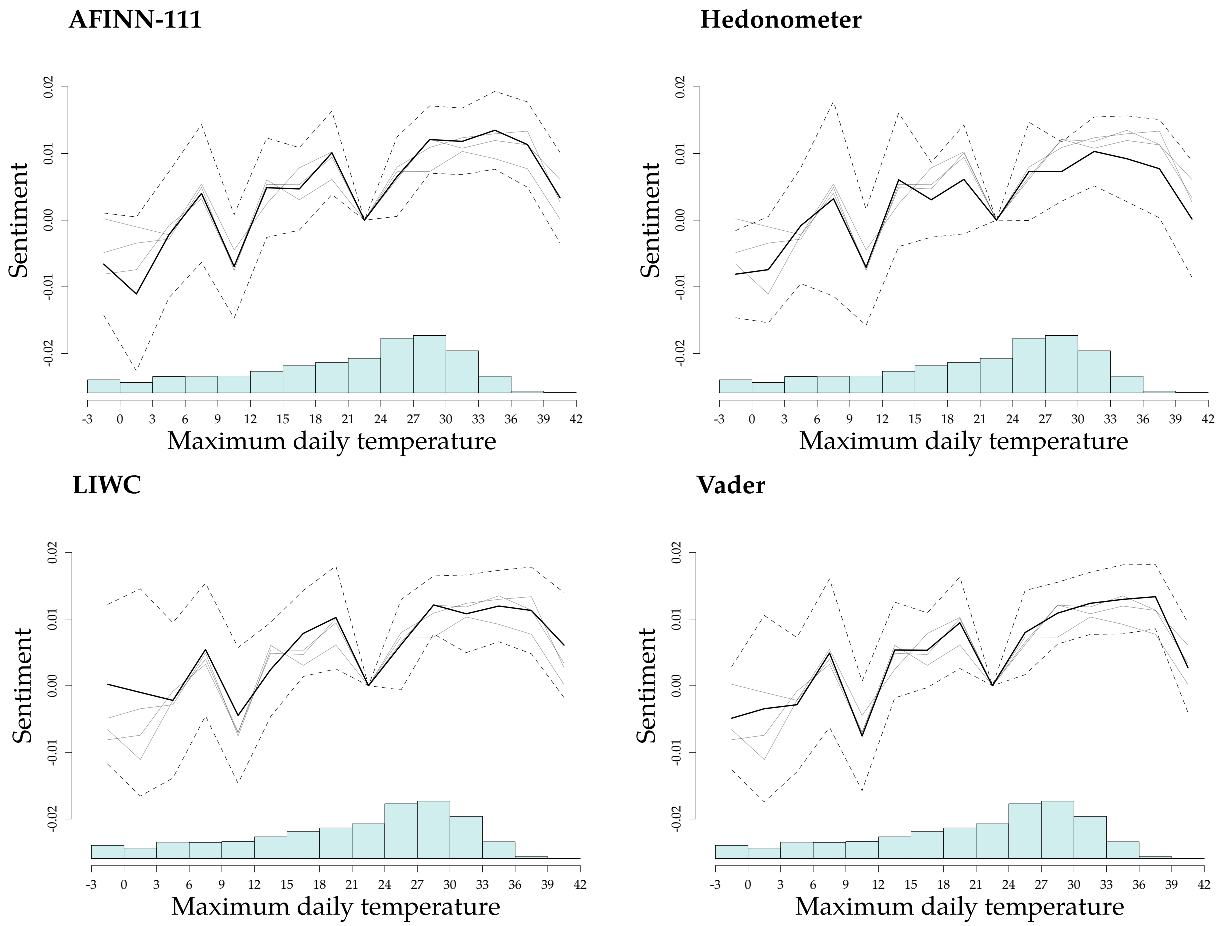

Figure A13: Monthly effect of temperature on Twitter sentiment

Notes: Monthly estimate of temperature on expressed sentiment. Outcome variable is standardized measures of expressed sentiment. Binned temperature variables indicate the number of days in a month in which the maximum daily temperature fell within the given range. Regressions also include precipitation, CBSA, month, and year fixed effects. Standard errors clustered by CBSA and date.

C.5 Residual variation

Figure A14 shows density estimates for columns of fixed effects in Table 4.

D Valuing expressed sentiment from parking and speeding tickets

To value the sentiment impact of receiving a ticket, I divide the sum of the average changes in sentiment on the seven days following the ticket by the median value of the stated ticket,

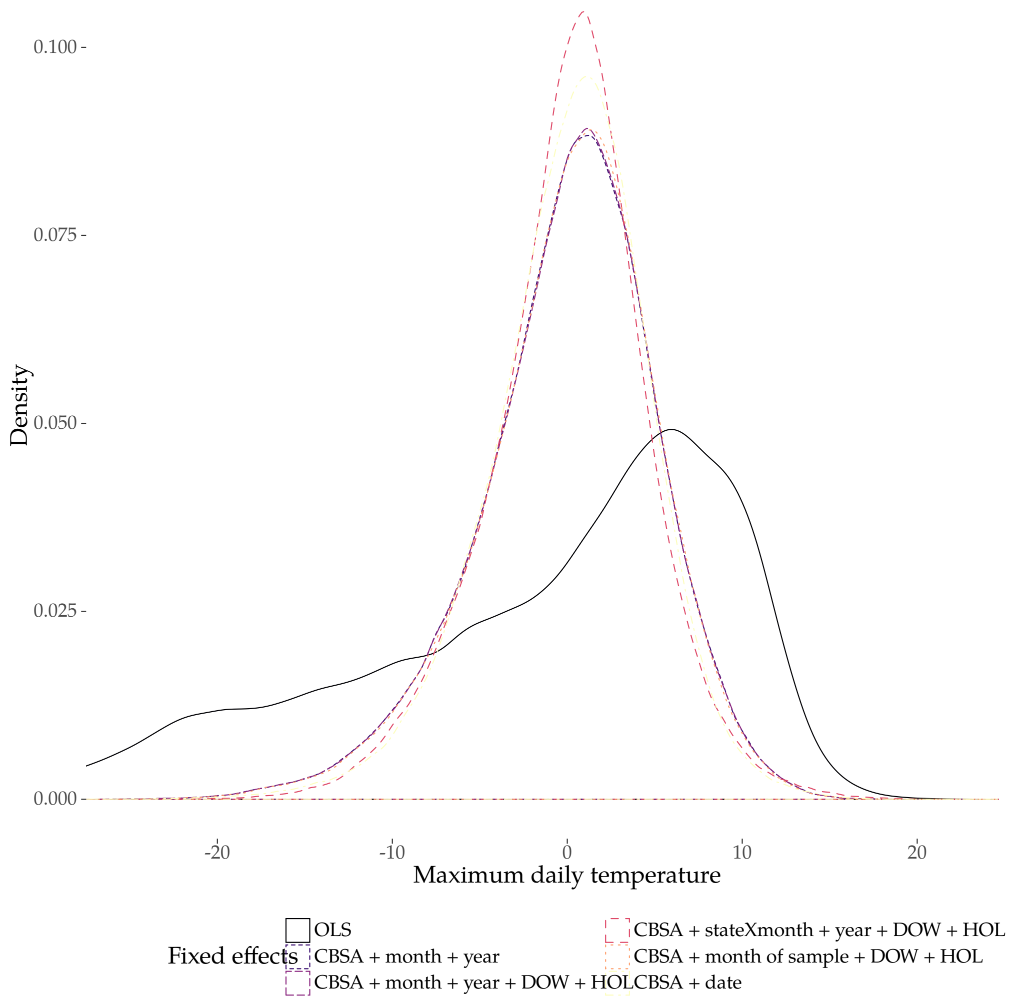

Figure A14: Residual variation

Notes: Density estimates of residual variation for columns of fixed effects in Table 4. Estimates are constructed by regressing maximum daily temperature on the given set of fixed of effects and plotting the density of the residuals from that regression. Each color and linetype represents a different combination of fixed effects.

$100.[13] Division of the $100 by 1.27 SD results in a per-SD value of $78.60. Table A13 applies this estimate to the contemporaneous estimates I obtain in column (1) of Table 4.

The estimates in column (3) of Table A13 imply, for example, that individuals in this sample would pay $0.79 to exchange a day with maximum temperature above between 35 and 40 C for a day with maximum temperature between 20 and 25 C. These results should be interpreted with some caution: this estimate is reliant both on the validity of both the empirical strategy estimating the effect of temperature on sentiment described earlier in

13. The mean in this sample is $164, driven by outliers with unrealistically large stated ticket costs. I use the median to mitigate the impact of these outliers and because it results in a more conservative estimate.

Table A13: Value of temperature

	(1)	(2)	(3)	(4)	(5)
Maximum daily temperature T					
$T \le 0$	−7.83	−8.98	−1.74	−11.73	−4.32
$T \in (0, 5]$	−8.19	−8.73	−2.95	−11.40	−4.94
$T \in (5, 10]$	−5.50	−4.98	−0.98	−7.08	−2.00
$T \in (10, 15]$	−0.77	−0.47	1.95	−1.24	1.20
$T \in (15, 20]$	−0.39	0.03	1.15	−0.50	0.35
$T \in (25, 30]$	0.65	0.93	−0.58	1.00	0.29
$T \in (30, 35]$	−2.60	−2.03	−3.36	−1.62	−3.63
$T \in (35, 40]$	−4.74	−4.71	−5.14	−3.60	−5.95
$T > 40$	−14.14	−12.82	−9.59	−11.90	−9.57
Fixed effects					
CBSA	Yes	Yes	Yes	Yes	Yes
Month	Yes	Yes			Yes
Year	Yes	Yes	Yes		Yes
DOW, Hol		Yes	Yes	Yes	Yes
S×M			Yes		
MOS				Yes	
State trends					Yes

Notes: Reproduction of Table 4, with coefficients multiplied by valuation of a unit SD change in standardized VADER sentiment. Value is obtained by dividing the median observed ticket cost ($100) by the total cumulative sentiment loss from a parking or speeding ticket (1.27 SD) in Figure 10.

the paper and the one described in this section. Because the estimate obtained $\sum_{k=0}^{K} \beta_k$ in Equation (3) serves as the denominator, if this strategy overestimates the impact of a parking ticket on sentiment, then the implied valuation of the changes in temperature will be underestimates, and vice versa.

D.1 Parking and speeding ticket dataset construction

To identify individuals who received parking tickets, I search for tweets containing all of the following words: "got", "a", and "ticket", and at least one of the following words: "parking", "speeding", or "traffic". After a manual review of these tweets, I remove entries which contain any of the following phrases: "out", "she", "mom", "dad", "almost", "got away with", "you", and "ya". This results in about 8,000 ticket-related tweets. Next, I search these tweets for "$" symbols followed by a number to identify the cost of these tickets. Finally, I identify the user responsible for each tweet and retrieve the full range set of their tweets in my sample. I limit the sample to users with at least 10 tweets within 30 days of ticket receipt.

E Climate projections

I project the annual amenity cost of rising temperatures across the United States on amenity value, measured in the change in SD of sentiment. I combine projected changes in climate with the estimates of the response of expressed sentiment to temperature. The nature of the projection exercise can be described mathematically as follows:

$$\int^{T} f(t)\Delta g(t)dt \tag{5}$$

where $f(\cdot)$ represents the damage function (valued in \$), such as the one estimated in Figure 3, and $\Delta g(\cdot)$ is the change in the distribution of climate. By integrating the product of f and g over the range of temperature T I obtain the total damages. Empirically, I estimate the shape of $f(\cdot)$ and combine climate and weather data to obtain $\Delta g(\cdot)$ in order to numerically approximate Equation (5).

With this framework, I conduct two exercises, referred to hereafter as the "baseline" and "adaptive" projection exercises. The baseline exercise projects damages using a single function for f, the estimate obtained by the splined model in Figure 4 and multiplied by the per-SD valuation from Section 5. The adaptive exercise uses the regionally-specific damage functions estimated in Figure 6 to project changes due to adaptation, also multiplied by the per-SD value. For each projection, I assume that CBSAs respond with the regional damage function that corresponds to the annual average temperature it currently experiences. In other words, if climate change warms an area to the extent that it moves from quartile 3 (the third-warmest region) into quartile 4 (the warmest region), then its damages are estimated using the quartile 4 damage function.

$g(\cdot)$ is estimated using the ensemble average from the output of 20 downscaled climate models[14], I compile average projections for each CBSA for the years 2006-2099. In order to de-bias the projections, I follow the prescriptions of Auffhammer et al. (2013) and add the difference between projected monthly decadal averages starting in 2026 and projected monthly averages from 2006-2025, then add those differences to the historical weather data from 2006-2015 to simulate future weather regimes for each decade while retaining historically observed variance in temperature. I estimate the difference in the distributions between baseline climate and the given future climate to obtain $\Delta g(\cdot)$.

Figure A15 documents the evolution of per-person annual damages over time, averaged over CBSAs and presented separately for RCP4.5 and RCP8.5, two different climate forcing scenarios of intermediate and high warming, respectively (IPCC 2014) and using both the baseline and adaptive methods described above. I estimate annual damages of increasing over time across all scenarios, with damages from RCP8.5 exceeding damages from RCP4.5. I find that the adaptive approach estimates larger damages than the baseline approach. Intuitively, it would seem that adaptation should mitigate the impact of climate change, but this does not occur here. The key element is that the damages from cold weather for the coldest areas of the United States are substantially smaller (see Figure 4) than they are for warmer areas. As a result, warming from climate change does not benefit those areas as much as it does in the baseline scenario. One interpretation is that since colder areas have already adapted to colder temperatures, climate change is less of an improvement than it would have been if those areas had not already adapted. This effect declines in importance as temperatures increase: the adaptive and baseline scenarios converge by end of century for RCP8.5.

The top panel in Figure A16 maps end of century damages under RCP8.5 by CBSA under the baseline scenario, documenting clear north-south heterogeneity in the extent of amenity costs due to climate change. The bottom panel does the same using the adaptive scenario.

14. Climate forcings drawn from a statistical downscaling of global climate model (GCM) data from the Coupled Model Intercomparison Project 5 (Taylor, Stouffer, and Meehl 2012) using the Multivariate Adaptive Constructed Analogs (MACA; Abatzoglou and Brown 2012) method with the Livneh (Livneh et al. 2013) observational dataset as training data.

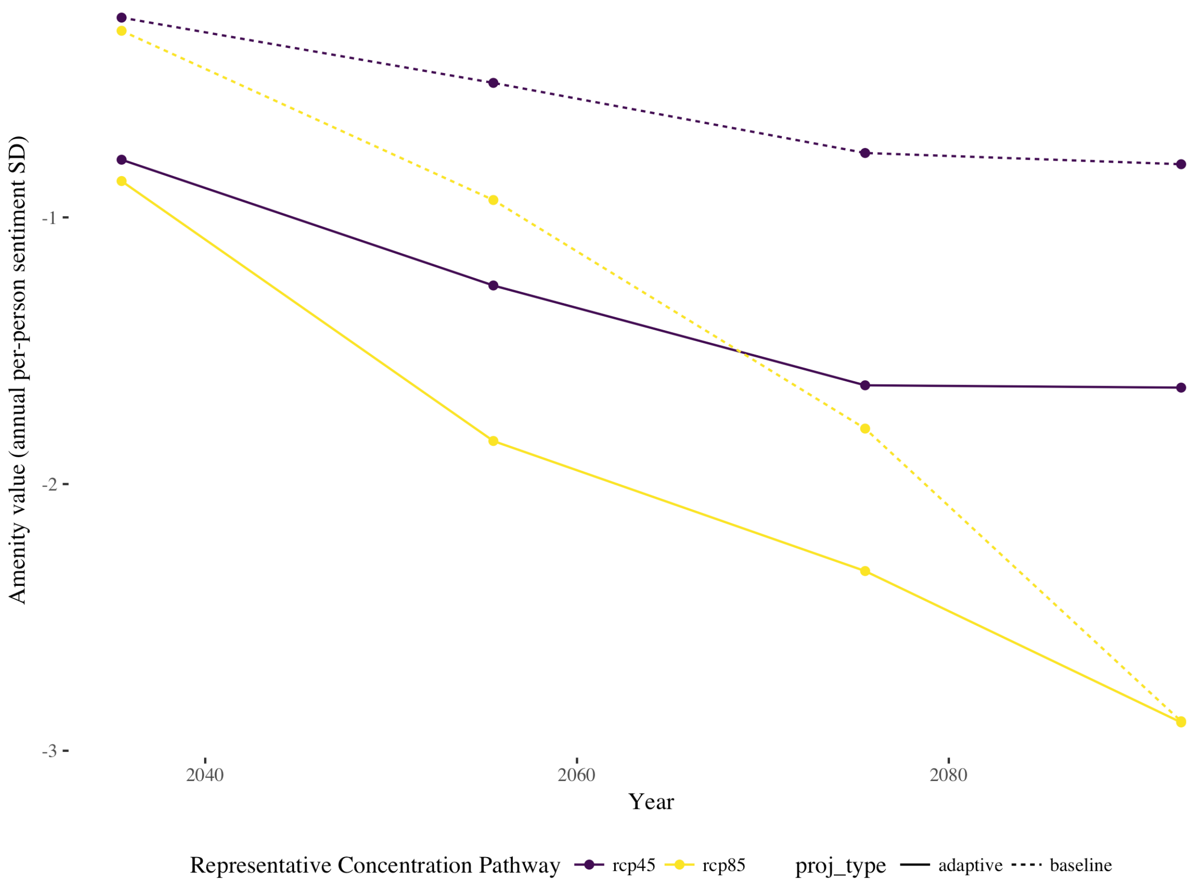

Figure A15: Projections of changes in amenity value over time

Notes: Projections of average change in amenity value over time, where average is the weighted average using total number of tweets per CBSA as the weights. Line color indicates the warming scenario, or Representative Concentration Pathway, used in the the projection data. Line type indicates whether the projection method was the baseline or adaptive method.

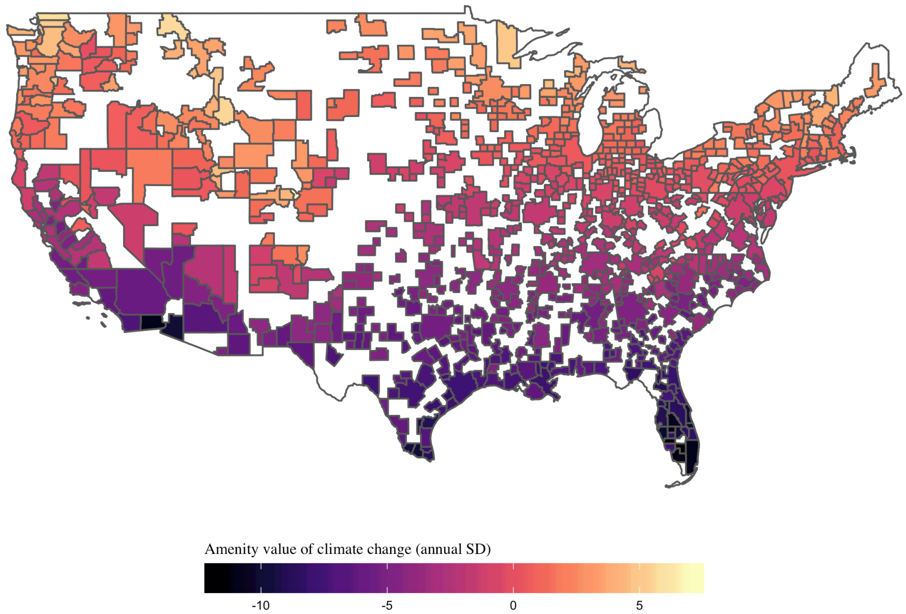

Figure A16: End of century projections of changes in amenity values (RCP8.5)

Notes: Top panel: projection of end-of-century climate damages in SD of expressed sentiment by CBSA under RCP8.5 using the baseline projection method. Bottom panel: projection of end-of-century climate damages in SD of expressed sentiment by CBSA under RCP8.5 using the adaptive projection method.

References

Abatzoglou, John T., and Timothy J. Brown. 2012. "A comparison of statistical downscaling methods suited for wildfire applications." *International Journal of Climatology* 32 (5): 772–780.

Auffhammer, Maximilian, Solomon Hsiang, Wolfram Schlenker, and Adam Sobel. 2013. "Using Weather Data and Climate Model Output in Economic Analyses of Climate Change." *Review of Environmental Economics and Policy* 7 (2): 181–198.

Bradley, Margaret Mm, and Pj Peter J Lang. 1999. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings.* Technical report. The Center for Research in Psychophysiology, University of Florida.

Dennisenn, J., Ligaya Butalid, Lars Penke, and Marcel A.G. Van Aken. 2008. "The effects of weather on daily mood: A multilevel approach." *Emotion* 8 (5): 662–667.

Dodds, Peter Sheridan, and Christopher M. Danforth. 2010. "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents." *Journal of Happiness Studies* 11 (4): 441–456.

Gilbert, CJ, and Eric Hutto. 2014. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In *Eighth International Conference on Weblogs and Social Media.*

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–161. arXiv: arXiv:1011.1669v3.

IPCC. 2014. *IPCC Fifth Assessment Report.* Technical report. Cambridge, United Kingdom and New York, NY, USA: IPCC.

Kahn, Jeffrey H., Renée M. Tobin, Audra E. Massey, and Jennifer A. Anderson. 2007. "Measuring emotional expression with the Linguistic Inquiry and Word Count." *American Journal of Psychology* 120 (2): 263–286.

Livneh, Ben, Eric A. Rosenberg, Chiyu Lin, Bart Nijssen, Vimal Mishra, Kostas M. Andreadis, Edwin P. Maurer, and Dennis P. Lettenmaier. 2013. "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions." *Journal of Climate* 26 (23): 9384–9392.

Nielsen, Finn Årup. 2011. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs." *arXiv preprint arXiv:1103.2903.*

Pennebaker, James W, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical report.

Robinson, W S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15 (3).

Tausczik, Y. R., and J. W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29 (1): 24–54.

Taylor, Karl E., Ronald J. Stouffer, and Gerald A. Meehl. 2012. *An overview of CMIP5 and the experiment design.* arXiv: arXiv:1011.1669v3.