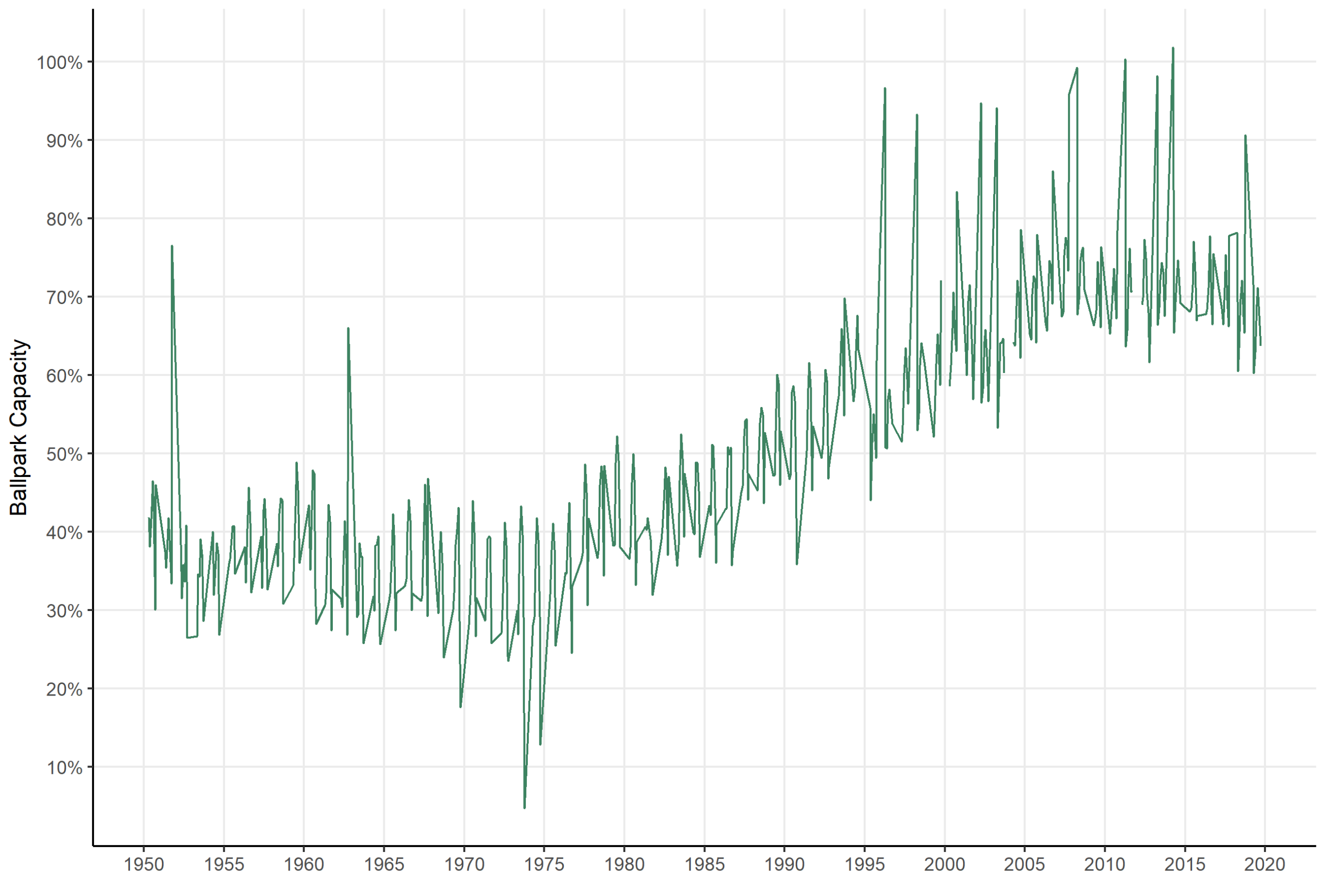
Online Appendix

The Willingness to Pay for a Cooler Day: Evidence from 50 Years of Major League Baseball Games

Kevin Kuruc   Melissa LoPalo   Sean O'Connor

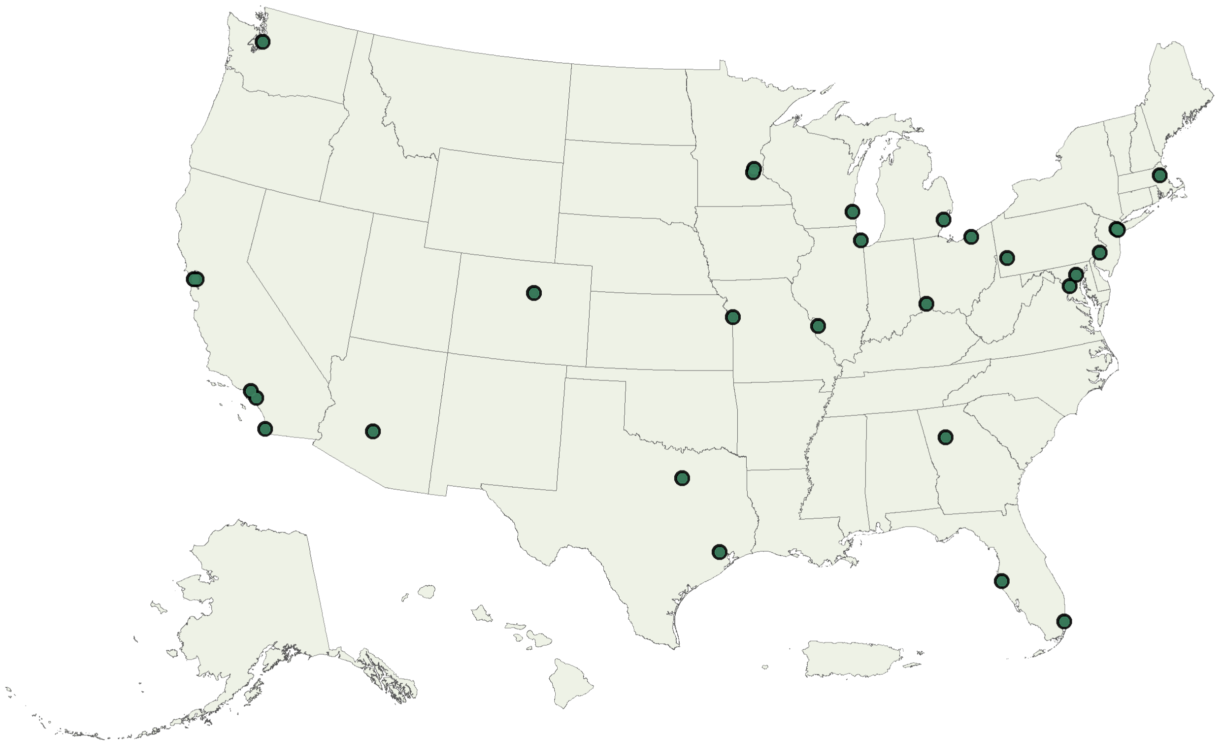
A Tables and Figures

Figure A1: Summary Statistics: Stadiums are More Full than in Past



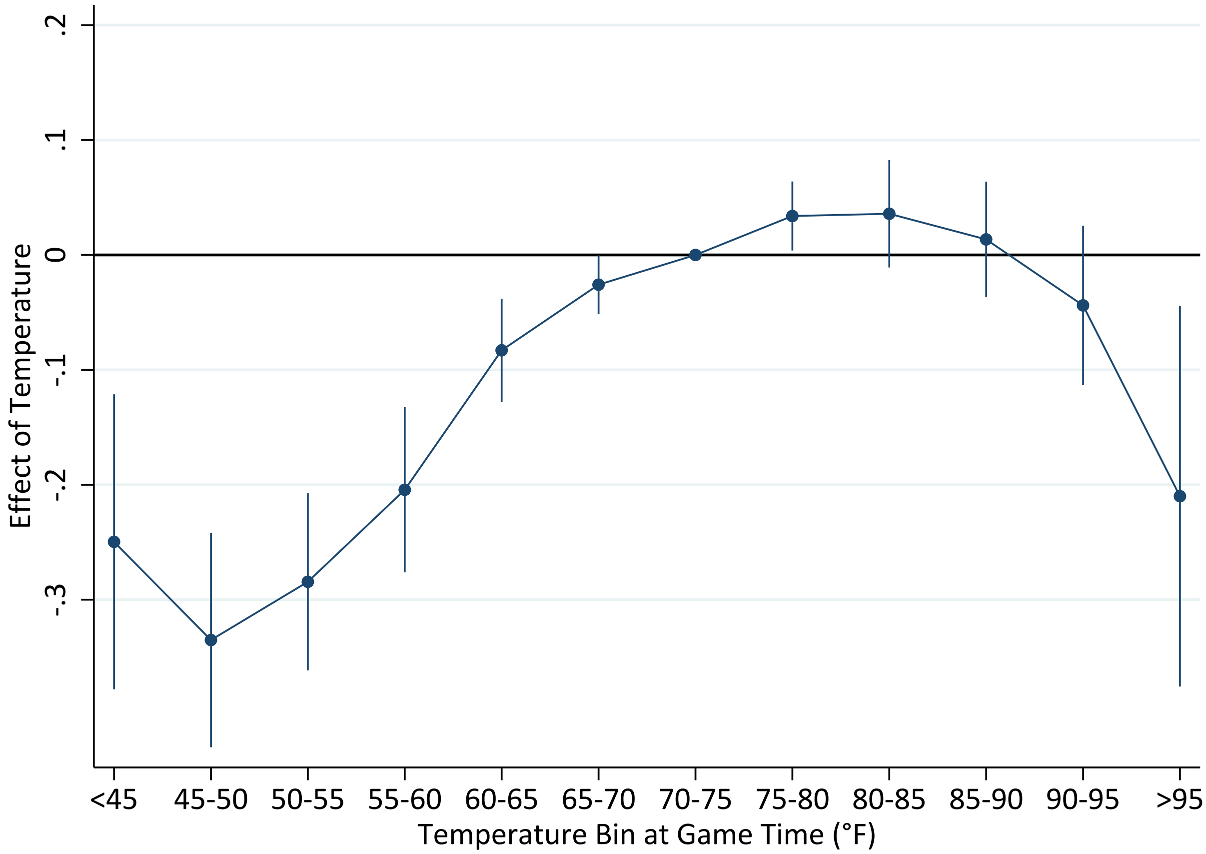
Note: This figure displays attendance as a percent of baseball stadium capacity for each month of each season from 1950 through 2019 using data from `Retrosheet.org`. The large spikes are generally associated with the first game of each season, which is the most well-attended game of the year for most teams.

Figure A2: CITIES WITH MLB STADIUMS



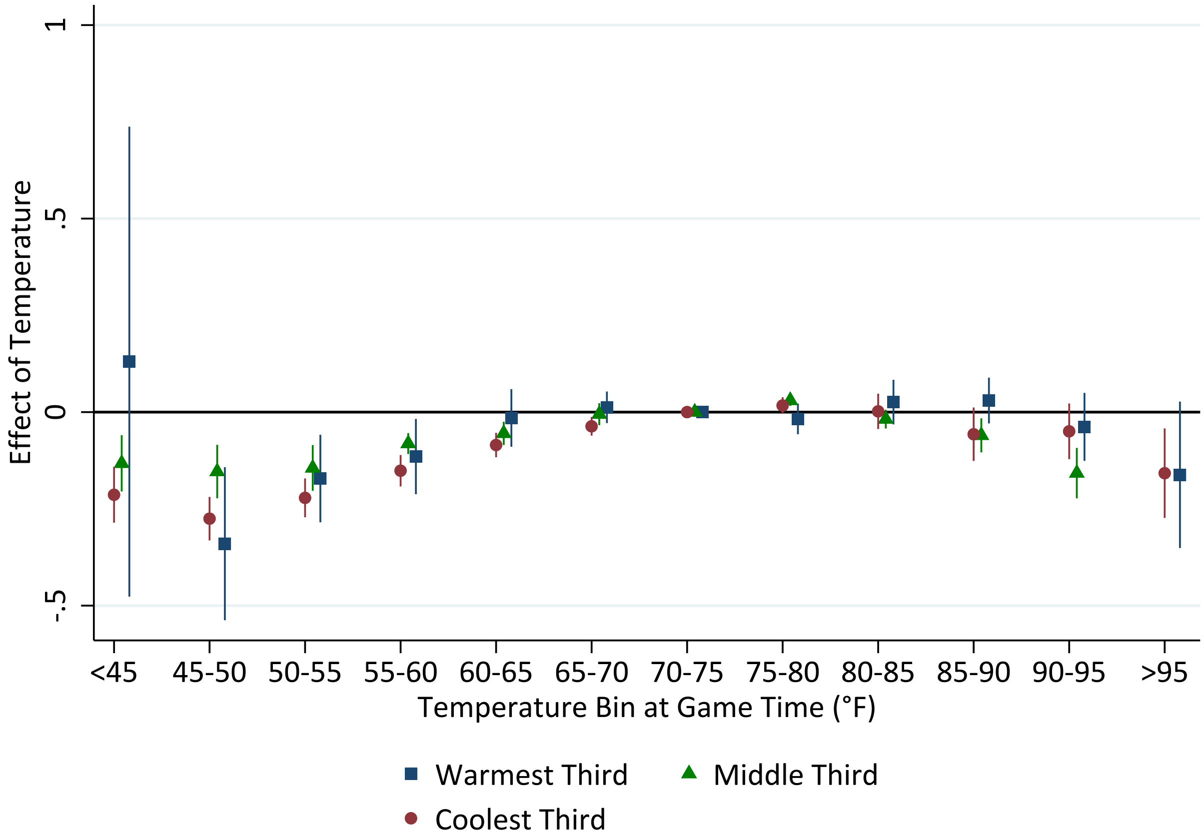
Note: This figure displays the location of the MLB stadiums in our sample. Some large cities, such as New York City and Chicago, have multiple stadiums operating at the same time. For a complete list of teams in the sample and their years of operation, see Table A1.

Figure A3: Estimates Robust to Limiting Sample to Teams Reporting Turnstile Attendance



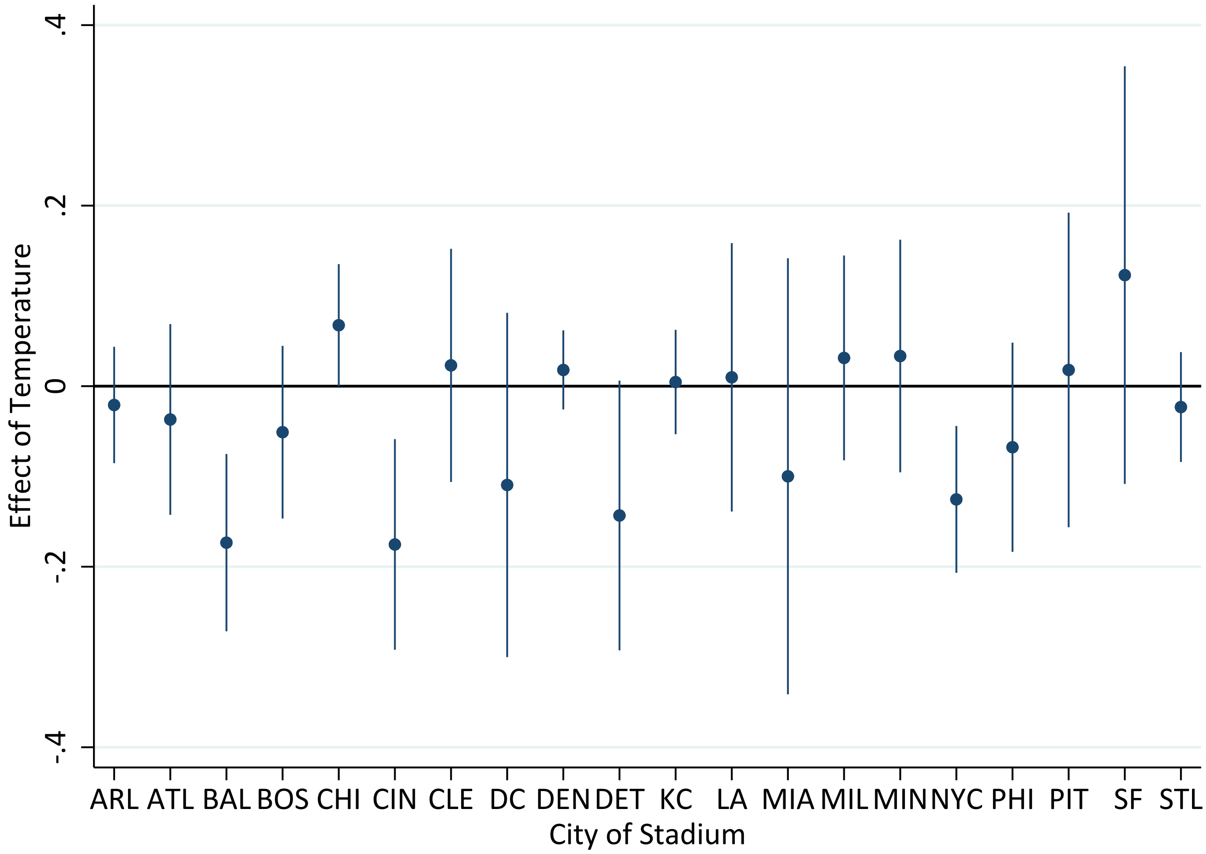
Note: This figure displays the results of Equation 1, run on the sample of National League games from 1950 to 1992. The outcome variable is game-level log attendance, and the independent variables of interest are indicators for game-time temperature (the average of hourly temperature readings for the three hours after the start of the game) falling into the temperature bin of interest. The regression also includes stadium by month fixed effects, visiting team fixed effects, month by year fixed effects, and controls for daily precipitation, the share of the last 100 games the home team won, indicators for day of the week, and an indicator for whether the game took place in the day or evening. Standard errors are clustered by stadium. Point estimates and 95% confidence intervals are shown.

Figure A4: NO EVIDENCE OF SYSTEMATIC DIFFERENCES BY USUAL CLIMATE



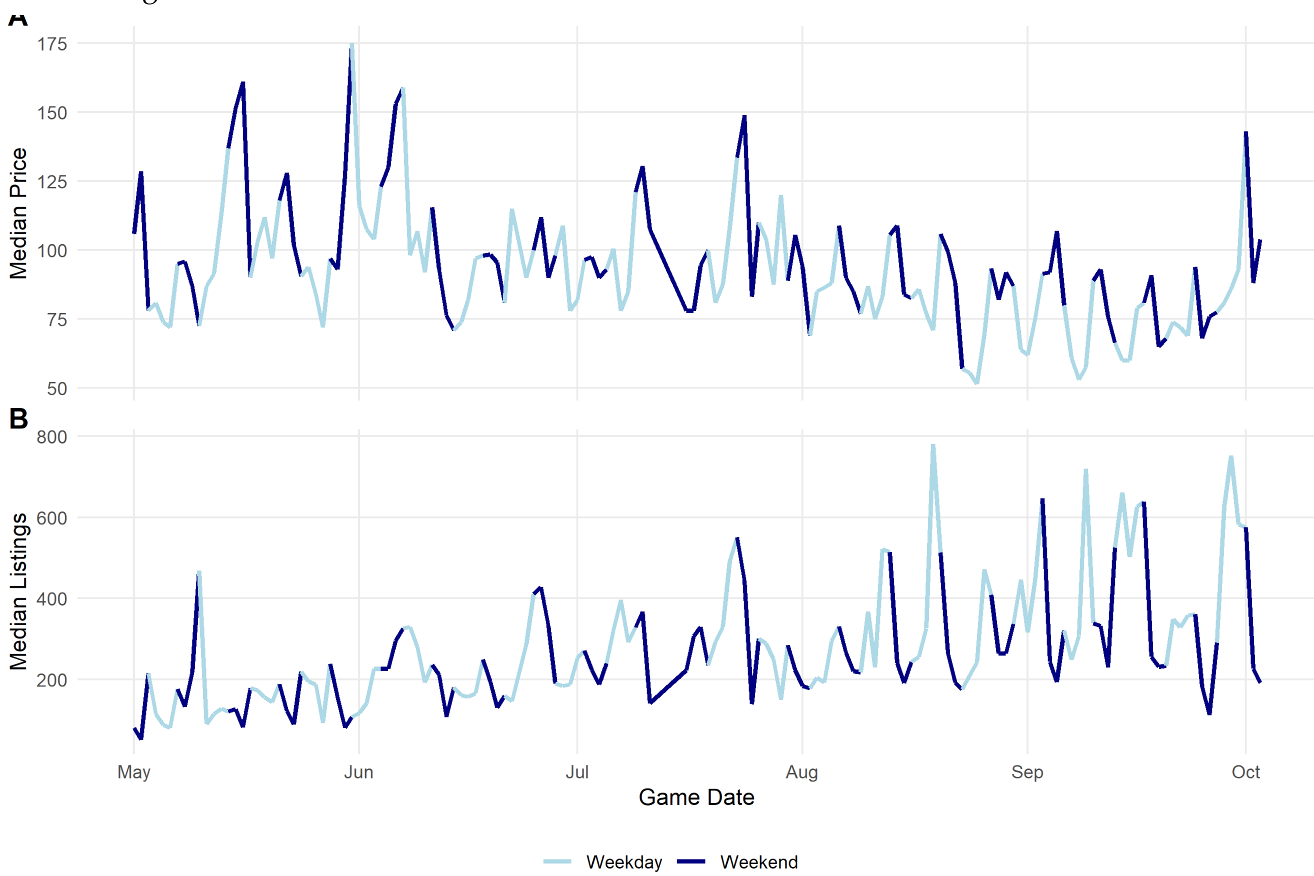
Note: This figure displays the results of Equation 1, run on the sample of MLB games from 1950 to 2000. The outcome variable is game-level log attendance, and the independent variables of interest are indicators for game-time temperature (the average of hourly temperature readings for the three hours after the start of the game) falling into the temperature bin of interest. These indicators are interacted flexibly with indicators for the stadium being in the warmest, coolest, or medium third of stadiums in the sample based on annual average temperature throughout the operation of each stadium. The regressions also include stadium by month fixed effects, visiting team fixed effects, month by year fixed effects, and controls for daily precipitation, the share of the last 100 games the home team won, indicators for day of the week, and an indicator for whether the game took place in the day or evening. Standard errors are clustered by stadium. Point estimates and 95% confidence intervals are shown.

Figure A5: LITTLE EVIDENCE OF HETEROGENEITY BY CITY



Note: This figure displays the results of Equation 1 on the sample of MLB games from 1950 to 2000, run separately for each city in the sample. The outcome variable is game-level log attendance, and the independent variables of interest are indicators for game-time temperature (the average of hourly temperature readings for the three hours after the start of the game) falling into the temperature bin of interest. Coefficients on indicators for game time temperature being above 90 degrees are displayed (the 90-95 and over 95 degree bins have been combined in this specification to increase statistical power). The regressions also include visiting team fixed effects, month by year fixed effects, and controls for daily precipitation, the share of the last 100 games the home team won, indicators for day of the week, and an indicator for whether the game took place in the day or evening. Point estimates and 95% confidence intervals are shown.

Figure A6: SEATGEEK SUMMARY STATISTICS: MEDIAN PRICE AND LISTINGS COUNT



Note: This figure displays median prices (Panel A) and median listing counts (Panel B) on Seatgeek from May through October of 2021, averaged by day of the season.

Table A1: TEAMS INCLUDED IN SAMPLE

	N. Home Games	First Year	Last Year
West			
Arizona Diamondbacks	1,780	1998	-
Colorado Rockies	2,146	1993	-
Los Angeles Angels	4,636	1961	-
Los Angeles Dodgers	4,872	1958	-
Oakland Athletics	4,029	1968	-
San Diego Padres	3,976	1969	-
San Francisco Giants	4,786	1958	-
Seattle Mariners	3,406	1977	-
Seattle Pilots	74	1969	1969
South			
Atlanta Braves	4,172	1966	-
Baltimore Orioles	4,922	1954	-
Houston Astros	4,578	1962	-
Miami Marlins	2,125	1993	-
Tampa Bay Devil Rays	1,778	1998	-
Texas Rangers	3,754	1972	-
Washington Nationals	1,207	2005	-
Washington Senators	734	1950	1960
Washington Senators	778	1961	1971
Northeast			
Boston Braves	193	1950	1952
Boston Red Sox	5,320	1950	-
Brooklyn Dodgers	532	1950	1957
New York Giants	520	1950	1957
New York Mets	4,341	1962	-
New York Yankees	5,171	1950	-
Philadelphia Athletics	304	1950	1954
Philadelphia Phillies	5,199	1950	-
Pittsburgh Pirates	5,200	1950	-
Midwest			
Chicago Cubs	5,238	1950	-
Chicago White Sox	5,137	1950	-
Cincinnati Reds	5,253	1950	-
Cleveland Guardians	5,122	1950	-
Detroit Tigers	5,282	1950	-
Kansas City Athletics	942	1955	1967
Kansas City Royals	3,944	1969	-
Milwaukee Braves	926	1953	1965
Milwaukee Brewers	3,867	1970	-
Minnesota Twins	4,608	1961	-
St. Louis Browns	251	1950	-
St. Louis Cardinals	5,382	1950	-

Table A2: Baseball Game Quality and Temperature

	(1) Duration	(2) Runs	(3) Home Runs	(4) Strike Outs	(5) Pitchers
<45	1.269	-1.017***	-0.692***	0.428**	-0.255**
	(1.084)	(0.269)	(0.113)	(0.139)	(0.088)
45-50	0.396	-0.872***	-0.558***	0.539***	-0.264***
	(0.860)	(0.212)	(0.080)	(0.123)	(0.064)
50-55	0.417	-0.656***	-0.416***	0.375***	-0.129**
	(0.604)	(0.164)	(0.062)	(0.105)	(0.047)
55-60	-0.465	-0.546***	-0.311***	0.228*	-0.157**
	(0.426)	(0.111)	(0.044)	(0.097)	(0.045)
60-65	-1.208**	-0.414***	-0.215***	0.128	-0.133***
	(0.352)	(0.085)	(0.034)	(0.069)	(0.030)
65-70	-0.566	-0.240**	-0.092***	0.117**	-0.111**
	(0.394)	(0.081)	(0.023)	(0.040)	(0.037)
75-80	1.122***	0.272***	0.116***	-0.154**	0.097***
	(0.309)	(0.073)	(0.023)	(0.048)	(0.023)
80-85	1.385**	0.404***	0.177***	-0.296***	0.127***
	(0.465)	(0.114)	(0.044)	(0.066)	(0.035)
85-90	2.831***	0.798***	0.305***	-0.276**	0.271***
	(0.550)	(0.171)	(0.059)	(0.089)	(0.048)
90-95	3.114**	0.881***	0.363***	-0.446***	0.262***
	(0.887)	(0.248)	(0.091)	(0.110)	(0.073)
>95	1.576	1.544***	0.513***	-0.253	0.457**
	(1.235)	(0.346)	(0.098)	(0.252)	(0.138)
Obs.	74,004	74,004	74,004	74,004	74,004
Mean of Dep.	161	9	2	11	5

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This figure displays the results of Equation 1, run on the sample of MLB games from 1950 to 2000. The dependent variables are indicators of the quality of game play, and the independent variables of interest are indicators for game-time temperature (the average of hourly temperature readings for the three hours after the start of the game) falling into the temperature bin of interest. The regressions also include stadium by month fixed effects, visiting team fixed effects, month by year fixed effects, and controls for daily precipitation, the share of the last 100 games the home team won, indicators for day of the week, and an indicator for whether the game took place in the day or evening. Standard errors are clustered by stadium.

Table A3: SeatGeek Price and Listing Summary Statistics

	Game Day			One Day Before			Two Days Before		
	Mean	P50	N	Mean	P50	N	Mean	P50	N
Average Price	445.70	90.50	1,992	278.23	94.00	1991	264.25	98.00	1980
Median Price	79.20	69.00	1,992	79.91	71.00	1991	81.13	73.00	1980
Lowest Price	32.52	25.00	1,992	30.18	24.00	1991	30.36	24.50	1980
Highest Price	2626.66	466.50	1,992	2755.85	590.00	1991	3232.35	683.50	1980
N. Listings	349.99	235.00	1,992	488.42	360.00	1991	537.72	402.50	1980

Note: This table presents summary statistics for SeatGeek ticket listing data from the 2021 MLB season. Data was accessed using SeatGeek's API, beginning on May 1, 2021 and ending on the final day of the season, October 3.

B Data Appendix

In this Appendix, we provide more information on the construction and contents of the weather, attendance, Seatgeek ticket price, and other data used in the main analysis.

BA Weather Data

NOAA's Integrated Surface Database We downloaded hourly temperature data from weather stations in the contiguous U.S. from the National Oceanic and Atmospheric Administration (NOAA's) Global Hourly - Integrated Surface Database (ISD). The ISD combines digitized weather station data from many data sources on over 20,000 stations into a single database, providing information on parameters such as temperature, wind speed, pressure, and precipitation at various time scales. We downloaded the ISD data using the worldmet package in R, and compiled a list of stations that reported continuously throughout the tenure of each baseball stadium in our sample. To ensure that a station reported continuously, we first remove any stations that stopped reporting for an entire year during the tenure of the stadium, and then further remove any stations that were missing more than 5% of 3-hourly temperature readings. These steps left 1,435 missing station-game observations in our weather dataset, which we fill in with the closest reporting weather station, if available. This results in a dataset with only 229 missing station-game observations, out of over 300,000 total. We then took the average of weather station temperature readings for the three hours starting at the start of each game, and then took the average for each weather station within 50km of the stadium, weighted by inverse distance. The worldmet package returns the weather data in UTC, so we convert to the local time zone before merging with the Retrosheet data.

Every stadium in the contiguous U.S. had at least one weather station nearby that reported continuously throughout its tenure. However, in many cases these stations reported only once per three hours for a period of time, usually in the late 1960's and early 1970's. In these cases, the station's identifying code often changed. A station's code consists of a USAF (U.S. Air Force) code and a WBAN (Weather-Bureau-Army-Navy) code. In cases where the code changed, one of the USAF or WBAN code always stayed the same, so we combine stations with the same USAF or WBAN code when judging the continuity of a station's reporting. NOAA has confirmed that these represent the same station, and we confirm that these stations are in the same location (within 1km in most cases, although it can be further and still be the same airport, for example). During the years that a station reported only once per three hours, if it's the only station available, the game-time assigned weather becomes the one reading in the three hour period, rather than an average of three readings.

Schlenker's Fine-Scaled Weather Data Set For the analysis of the effects of weather on attendance, we source our daily precipitation data from Wolfram Schlenker's Daily Weather for Contiguous United States dataset, which is derived from the PRISM dataset. This dataset provides total precipitation at the 2.5 mile resolution. This dataset creates a balanced panel of monitoring stations by interpolating missing station observations with the distance-weighted average of the cumulative density function of surrounding stations. This dataset is available from 1900-2019.

PRISM Data For the analysis of the effects of weather on Seatgeek ticket prices, we use the PRISM weather dataset, developed by Oregon State University's PRISM Climate Group, for our daily precipitation data. These data are compiled from monitoring stations across the U.S. and are spatially interpolated to a 4km resolution using the PRISM model, which accounts especially for the effects of elevation on the spatial distribution of temperature and precipitation. The dataset provides daily total precipitation, which is the sum of rain and melted snow for the day. We download PRISM data using the prism package in R.

Weather Projections from CMIP6 We use the average of the 35 global climate models (GCM's) from the Coupled Model Intercomparison Model 6 (CMIP6). The data are downloaded from the NASA Earth Exchange (NEX) Global Daily Downscaled Projections (GDDP). The NEX-GDDP is downscaled to improve the resolution of output from the GCM's, allowing for analysis at the 0.25 degree latitude/longitude resolution. Daily maximum and minimum near surface air temperatures are provided at the daily level, which we average to form a measure of daily average temperature. We then calculate a count of days between 85-89.5 degrees, which corresponds to the 25th and 75th percentile for game-time temperatures over 95 in our sample, for each grid cell for each year from 2080-2090, and then create a decadal average of the yearly count. We calculate these averages for two climate change scenarios from the new scenario framework introduced in the IPCC Sixth Assessment Report. SSP5-8.5 is a high-emissions scenario with no mitigation policy, in which economic growth is fueled by fossil fuels. This scenario produces warming of 4.4 degrees Celsius by end of century. SSP2-4.5 is a middle-of-the-road scenario in which emissions begin to fall by the mid 21st century, and socioeconomic factors are stable. This scenario produces warming of 2.7 degrees Celsius by 2100.

BB Baseball Data

Retrosheet Data We download data on game-level attendance and game play from retrosheet.org using the retrosheet package in R. We make a few adjustments to the retrosheet data to derive a consistent set of team-stadium observations. First, we code any team that changed its name but remained in the same city to be a single team.

- The Florida Marlins became the Miami Marlins; we code them as the same team.
- The California Angels became the Anaheim Angels in 1996 and subsequently the Los Angeles Angels of Anaheim in 2005; we code them as the same team.

If a team moved cities and changed names, we consider the team in the new city to be a new team.

- The Kansas City Athletics moved to Oakland and became the Oakland Athletics in 1968; we code these as different teams.
- The Washington Senators moved to Minnesota and became the Minnesota Twins in 1961. A new Washington Senators team sprung up in its place, from 1960-1971, before moving to Dallas-Fort Worth and becoming the Texas Rangers in 1972. The original Washington Senators, replacement Washington Senators, Minnesota Twins, and Texas Rangers are each coded as different teams in our sample.
- The Seattle Pilots moved to Milwaukee and became the Milwaukee Brewers in 1970; they are coded as different teams.

We combine a list of stadiums from Retrosheet with park configuration data from Seamheads. Seamheads.com provides data on capacity and type of cover for each stadium by year. We use this information in combination with Retrosheet to form our measure of attendance as a share of capacity as well as for our sample restriction to only open-air stadiums. We drop some one-off venues from our dataset: the Oakland Athletics played 6 games in Las Vegas in 1996, the Atlanta Braves played 1 game at Fort Bragg in 2016, the Tampa Bay Rays played 3 games at the Ballpark at Disney's Wide World in 2007, and the Pittsburgh Pirates played 1 game at BB&T Ballpark at Bowman Field in 2017. Our results are robust to excluding or including these venues from the sample.

Seatgeek Data We accessed the SeatGeek.com API on a daily basis at 8:00A.M. Central Time starting on May 1, 2021 and ending on October 3, 2021, the final day of the 2021 regular season. Each access of the API collected data on that day and the following two days' MLB games. The variables collected include:

- Home and away teams
- Date of the API request
- Game date
- Game start time
- Geographic information, including the city, state, latitude, and longitude of where the game was to take place
- The number of ticket listings
- Mean, median, minimum, and maximum ticket price for seats available during the game

The data are at the game-date level, not the individual ticket level. It should also be noted that the number of ticket listings does not equal the number of tickets available; for example, a listing for four seats next to each other would be counted as one listing.

BC Other Data

American Time Use Survey Data We use American Time Use Survey (ATUS) to calculate approximately how much time the average American spends outside at hot and cold temperatures. To create this statistic, we download the 2003-2018 multi-year ATUS data and merge it with daily average temperature data from Wolfram Schlenker's version of the PRISM data, by county and date of the diary entry. We then classify activities as taking place outdoors vs. indoors to calculate how much time is spent outside each day. We follow Graff Zivin and Neidell (2014) in defining time outdoors as time spent outdoors away from the home, or doing specific home activities such as walking, exterior maintenance, and lawn, garden, and houseplants. Outdoor labor time is not included in the main measure, as the American Time Use Survey does not distinguish whether the "respondent's workplace" is indoors or outdoors. However, in a secondary estimate we follow Graff Zivin and Neidell (2014) in defining a set of climate-exposed industries, where much of the work is likely done exposed to outdoor temperatures: agriculture, forestry, fishing, and hunting; mining; construction; manufacturing; and transportation and utilities industries. We then count time at work in these industries as additional time outdoors.

References

Graff Zivin, Joshua and Matthew Neidell (2014) “Temperature and the Allocation of Time: Implications for Climate Change,” *Journal of Labor Economics*, 32 (1), 1–26.