Effect of Temperature and Weather Shocks on Health at Birth: Evidence from the US

Johanna Kutz*

October 9, 2023

Preliminary and incomplete - please do not cite or circulate

Abstract

Understanding in-utero exposure to extreme weather events is key to mitigating climate change's impact on health at birth. Using detailed historic weather records and data on infants born in the US between 1989-2004, we investigate how in-utero exposure to weather events, such as heat and cold waves or rainfall, impacts infant's health at birth. We focus on the effects of heat shocks on birth outcomes and systematically investigate heterogeneity therein using the causal forest, a recently developed causal machine learning technique. Exposure to a heat shock significantly reduces birth weight by around 6 grams on average and increases the small for gestational age (SGA) birth rate. We find substantial heterogeneity in the effect of heat shock exposure on birth weight. Especially infants born to black, Mexican, or low-educated mothers are disproportionately prone to health risks from extreme heat exposure.

JEL classification: C21, I14, Q54

Keywords: Climate Change, Extreme Weather, Health at Birth, Causal Machine Learning, Heterogeneity

*University of St. Gallen, johanna.kutz@unisg.ch

1 Introduction

Climate change is the biggest health threat humanity faces in the 21st century (World Health Organization, 2018). In recent years, extreme weather events such as heat waves, cold spells, and heavy rainfall have become increasingly frequent and severe due to climate change. Especially vulnerable groups, including older populations, women, and children are disproportionately prone to climate-sensitive health risks. While the impact of in-utero exposure to these events on health has been extensively studied, less attention has been paid to possible heterogeneity in these effects and the anaysis of most vulnerable groups. Pregnant women are considered a particularly vulnerable group when it comes to exposure to extreme weather events due to the changes in their bodies physiology (Samuels et al., 2022). It is crucial to understand how exposure to weather in utero affects the fetus, as it may affect the infant's long-term outcomes, in line with the fetal origins hypothesis[1].

This paper investigates the effects of in-utero exposure to several extreme weather events in the US between 1989 and 2004 and systematically explores potential heterogeneity in these effects. Weather events considered include heat and cold events, and rainfall shocks. We analyze their effects on multiple outcomes of health at birth, like standardized birth weight, small for gestational age birth (SGA)[2], gestation length, and 5-minute Apgar score[3]. Following the literature, we first analyze the effects of these shocks on the birth outcomes of interest using a fixed effect regression model, counting the days of exposure in certain temperature bins. Our main interest, however, lies in understanding whether there is heterogeneity in the effect of extreme heat on health at birth. Therefore, we study the conditional average treatment effect (CATE) of heat shocks on health at birth, which we estimate using a causal forest (Athey et al., 2019). We uncover heterogeneity in the effect and describe the most vulnerable groups and attempt to learn about possible mechanisms underlying the effect of heat shocks on health at birth.

We show that in-utero exposure to extreme heat events generally has adverse effects on infants' health at birth, where heat shocks can reduce birth weight by up to 6 grams. We define the shock as the average temperature of 5 consecutive days exceeding at least 30 °C and the temperature of the 9th hottest day per county. These heat shocks do particularly affect the most vulnerable infants, as it significantly increases the SGA rate. Analysis of heterogeneity in the effects reveals that especially infants born to black, Mexican, and low-educated mothers are disproportionately prone to heat-related health issues. Our decomposition, however, reveals that none of these factors on their own can explain the heterogeneity found.

Analysis of the entire temperature distribution reveals that especially exposure to hot temperatures shows the strongest negative effects on weight-related birth outcomes, while exposure to cold temperatures tends to have mitigating effects. Exposure to extreme rainfall does not significantly affect any of the measures of health at birth used. Exposure to a day with temperatures of more than 32° C decreases birth weight by

[1] The fetal origins hypothesis posits that conditions and shocks in utero can have long-lasting impacts on human capital, by influencing, for example, health, educational attainment, and adult earnings. See for example Almond and Currie (2011) for a review of this literature.

[2] A birth weight of less than 10th percentile for gestational age is considered small for gestational age.

[3] The Apgar Score is a key measure of infant health right after birth, focusing on activity and respiration of the newborn (American College of Obstetricians and Gynecologists., 2015).

0.57 grams, compared to a day in a temperature range of $0-24°$ C. Exposure to cold days shows smaller, yet significant increases in standardized birth weight. Especially the second and third trimesters show a strong response to extreme temperature exposure, whereas the first trimester does not seem very vulnerable.

Even though the effects of extreme weather events on health has found much attention, heterogeneity in these effects is not well understood. So far, analysis of heterogeneity has mainly been limited to infant's sex, mother's race, and mother's education (Chen et al., 2020, Deschênes et al., 2009, Le and Nguyen, 2021). Understanding the heterogeneity in the effects of extreme weather shocks on health at birth is crucial. It can help in identifying the most vulnerable groups and provide insight into the underlying mechanisms through which these shocks affect fetal development and health outcomes at birth. This eventually leads to improved strategies and targeted interventions to prevent damage from extreme weather events.

The analysis of heterogeneity in the effects of heat shocks on health at birth reveals strong heterogeneity, where the most negative impact is concentrated among black, Mexican, and low-educated mothers. Since the analysis of the average effect using counts of days in temperature bins showed the largest effects of exposure to extreme heat, we will focus on heat shocks for the heterogeneity analysis. We estimate the conditional average treatment effect of a temperature shock on birth outcomes of interest using a causal forest (Athey et al., 2019). To gain an understanding of how different maternal characteristics might affect treatment outcomes, the study compares the average characteristics of the 10% most and 10% least affected groups. This analysis reveals that the most affected group comprises primarily black, Mexican, and low-educated mothers, suggesting that they may have limited access to protective measures against heat exposure. Additionally, we employ a newly developed decomposition approach to uncover drivers of heterogeneity. Using this decomposition, we are able to isolate the effects of a single variable, while keeping the others comparable. Our analysis suggests that mother's race and Hispanic origin alone are not significant drivers of heterogeneity. This observation may be due to the fact that these variables serve as proxies for more complex factors such as living conditions and socioeconomic status, which could be the true drivers of heterogeneity.

A large body of literature analyzes how extreme weather events affect health and well-being. Focusing on mortality rates and hospital admissions Karlsson and Ziebarth (2018) investigate effects of high ambient temperature in Germany. They find that extreme heat immediately increases hospitalizations and deaths. Similarly, Deschênes and Greenstone (2011) analyze temperature effects on mortality in the US, finding that extremes on both ends of the temperature scale lead to increased mortality rates. Following the fetal origins hypothesis by David J. Barker, a growing part of the economic literature analyzes how in utero exposure to extreme weather events affects later life outcomes (Wilde et al., 2017, Isen et al., 2017, Chang et al., 2022). Evidence on the effect of temperature exposure on later life outcomes is mixed. Wilde et al. (2017) find exposure to increased temperature leads to higher educational attainment and literacy, whereas Isen et al. (2017) find such exposure leads to reduced adult earnings.

Most closely related to this study is a strand of literature that analyzes how extreme weather events during pregnancy affect birth outcomes. Studies focus on different countries of interest, like the US (Currie and Rossin-Slater, 2013, Deschênes et al., 2009), China (Chen et al., 2020), Vietnam (Le and Nguyen, 2021), and sub-Saharan Africa (Bratti et al., 2021). These studies use different ways to measure the exposure to extreme temperature events, i.e., Deschênes et al. (2009), Chen et al. (2020) use the count of days in certain temperature bins, Andalón et al. (2016), Le and Nguyen (2021), Molina and Saldarriaga (2017) use deviations from the long-term mean. All studies share the same conclusion: extreme weather events can have detrimental effects on health at birth. Especially for birth weight, studies find significant and robust reductions in birth weight and increases in low birth rates caused by extreme heat events. Similar to this study, Deschênes et al. (2009) analyze the effects of extreme heat on birth weight and low birth weight in the US for births between 1972-1988. They find that exposure to extreme heat events during pregnancy reduces birth weight and increases the prevalence of low birth weight (LBW) birth. Considering climate change predictions, the overall reductions correspond to losses of 7.5 to 11.5 grams. The effect is especially profound in the second and third trimesters, which can explain up to 95% of the effect found. Overall, the literature lacks a good understanding of possible heterogeneity in the effects of temperature on health at birth. Heterogeneity in the effects has mainly been studied with regard to mother's race, mother's education, and sex of child (Deschênes et al., 2009, Le and Nguyen, 2021).

While a large body of literature assesses the effects of in utero exposure to weather events (Andalón et al., 2016, Deschênes et al., 2009, Rocha and Soares, 2015, Zhang et al., 2020), they mostly neglect a fundamental problem regarding a mechanical correlation between gestation length and probability of exposure. Only Currie and Rossin-Slater (2013) discuss this problem when analyzing the effect of hurricane exposure on birth outcomes. Measuring temperature exposure during the gestational period leads to a problem since a longer gestation length increases the probability of exposure to an extreme event. This simultaneous interaction between the exposure and the mediator results in challenges with the identification of the causal effect of interest and in most cases leads to biased estimates if not handled carefully. To eliminate the simultaneity of temperature exposure and gestation length, the literature mostly imposes full 9 months of pregnancy for each birth. These fixed pregnancy lengths are either counted backward from the date of birth or forward from the date of conception. Each of these strategies leads to biases in the effect estimates, which in most cases are not discussed further. To overcome this problem, we analyze the effects on outcomes standardized for gestational age.

This paper contributes to the literature on in-utero exposure to weather shocks in three ways. First, it marks a significant contribution to the literature by introducing cutting-edge machine learning techniques to uncover heterogeneity in the effect of environmental factors on infant health. The paper highlights the potential of machine learning approaches for advancing our knowledge in the field of environmental health research. By leveraging recent causal machine learning techniques, we can identify the drivers of heterogeneity and characterize the most vulnerable groups, which may offer valuable insights for policymakers.

Second, this study is the first to systematically explore heterogeneity in the effect of in-utero exposure to heat waves on infant health. To the best of our knowledge, heterogeneity has been studied only by mother's race, mother's education, and sex of child (Deschênes et al., 2009, Le and Nguyen, 2021). This paper addresses an important gap in our understanding of the complex relationship between heat exposure during pregnancy and infant health outcomes. Exploiting heterogeneity by several mothers' characteristics allows us to understand the effects of extreme weather events beyond averages and helps to identify patterns of vulnerability.

Additionally, we provide new evidence for a range of birth outcomes, which control for differences in gestation length, and discuss problems arising from the mechanical correlation of gestation length and shock exposure. Only a few of the aforementioned studies consider outcomes other than birth weight and gestation length (Andalón et al., 2016, Molina and Saldarriaga, 2017). We additionally estimate the effect on standardized birth weight, standardized 5-minute Apgar score, and small for gestational age birth. We highlight the mechanical correlation between gestation length and the exposure measure and discuss issues arising from solutions usually employed in the literature.

The remainder of the paper is structured as follows: The next section provides an overview of the birth data and weather data sources. We then describe the empirical strategy and go on to describe the results. The last section discusses the results and concludes with a discussion of possible implications of the findings.

2 Data

2.1 Birth Data

Microdata on births in the US between 1989 and 2004 are taken from the National Vital Statistics System of the NCHS (National Center for Health Statistics, 1989-2004). The data provides information on births in the United States, based on information abstracted from birth certificates. The public use files contain information on the mother's county of residence only for counties with a population of at least 100,000 residents. The analysis therefore mainly focuses on highly populated areas on the US coasts, including a total of 525 distinct counties. See figure A.1 for an overview of counties considered. The data contains information on the socioeconomic and demographic background of the mothers, such as race, age, educational attainment, marital status, childbearing history, prenatal care, and information on medical risk factors. Further, it contains health information on the infant such as sex, gestational age, birth weight, Apgar score, plurality, and month of birth.

Details for trimester dates are calculated using the mother's last reported menses and gestation length in weeks. Since the exact date of birth is not given in the data, we need to approximate pregnancy start, trimester borders, and end dates. To determine the start of the pregnancy, we count forward from the mother's last reported menses. As Currie and Rossin-Slater (2013) point out, counting backward from the date of birth shifts trimester borders and therefore induces bias. We set the start date of the pregnancy as the month after mother's last reported menses. From this, we set the first trimester as the following three months, and the second trimester as the three months thereafter. The last trimester is determined by the month of birth and the gestation length. We, therefore, allow for differing lengths in the third trimester depending on the gestation length.

To measure infants' health at birth, we use several proxy measures. A widely used proxy measure for health at birth is birth weight. Especially low birth weight is strongly associated with adverse health outcomes Almond et al. (2005). But given the mechanical correlation between gestation length and exposure to extreme weather events, we mainly concentrate on outcomes, that account for differences in gestation length. Our main outcomes are therefore standardized birth weight[4], standardized 5-minute Apgar score and a measure for small-for-gestational age (SGA) birth. Additionally, we consider gestation length but concentrate on effects in the first and second trimesters.

We focus our analysis exclusively on singletons born in the United States. The rationale behind this decision is that multiple pregnancies tend to carry greater medical risks and complications for both the mother and the children involved (Kramer, 1987). We also remove the 1% most extreme outliers at both ends of the distribution for birth weight and gestation length. Moreover, we exclude all observations with missing data in the relevant variables of interest. Table 1 provides an overview of the birth data that we consider. Our sample comprises information on $19,865,677$ pregnancies that occurred in the US between 1989 and 2004. On average, the infants in our sample have a birth weight of 3401.86 grams and are born after 39.11 weeks of gestation. Approximately 15% of the births are preterm, and 4% are low birth weight. The average age of mothers in our sample is 28.22 years, and most of them are married and white. More than half of the mothers attended college, while 11% smoked during pregnancy.

2.2 Weather Data

Meteorological data were obtained from two sources: the North America Land Data Assimilation System (NLDAS) available on CDC WONDER (`https://wonder.cdc.gov/`) and the National Centers for Environmental Information (NCEI, `https://www.ncei.noaa.gov/`). We gather information on ambient temperature, rainfall and sunshine exposure from the North America Land Data Assimilation System, which can be accessed on county level[5] from CDC WONDER. Information on snowfall and snow density is obtained from NCEI for single weather stations. For required aggregation on the county level, we average data from all weather stations in the corresponding county. Information on snowfall is not recorded for all days in some counties. Therefore, we miss information on snowfall for 1,377,089 observed pregnancies.

We use measures for average ambient temperature and information on rainfall (in mm) to describe our main (extreme) weather events. We additionally gather monthly averages for each county and year for temperature, rainfall, snowfall, and sunshine (in KJ/m^2). Several studies have found effects of either of these weather conditions on health or health at birth (i.e., Andalón et al., 2016, Rocha and Soares, 2015, Zhang et al., 2020).

[4] Standardized birth weight is calculated as follows: First, we calculate mean reference birth weight and its standard deviation for given gestation length and sex. The reference population contains singletons, non-smokers, and pregnancies without complications. We standardize each birth weight by subtracting the reference weight given individuals gestation length and sex and dividing by the corresponding standard deviation. The standardized birth weight now fully controls for gestation effects and possible gender effects. We use the same procedure for the standardized 5-minute Apgar score.

[5] The North America Land Data Assimilation System provides the following information on county aggregation: The county coded represents the spatial average of data observations from 14x14 kilometer square (1/8 degree) geographic-area grids. Grids are coded to the county that includes the grid centroid. For small counties where no grid centroid fell within county boundaries, county data are aggregated from grids where most of the grid area fell within county boundaries.

Table 1: Summary Statistics for Birth Data

Variables	Mean	SD
Infant Characteristics		
Birth Weight in grams	3401.86	482.59
Standardized Birth Weight	-0.04	1.02
Low Birth Weight	0.04	0.19
Small for Gestational Age	0.11	0.31
5-min. APGAR Score	8.97	0.55
Gestation Length	39.11	1.74
Preterm Birth	0.15	0.36
Assisted Ventilation	0.02	0.13
Male	0.51	0.50
Mothers Characteristics		
Mother's Age	28.22	5.73
Married	0.81	0.39
Race - White	0.82	0.38
Race - Black	0.13	0.34
Race - Other	0.05	0.21
Non-Hispanic	0.88	0.32
Hispanic - Mexican	0.05	0.23
Hispanic - Other	0.06	0.24
Education - <HS	0.03	0.18
Education - Highschool	0.10	0.30
Education - some College	0.55	0.50
Education - College +	0.32	0.47
Smoking during Pregnancy	0.11	0.31
Weight Gain in pounds	30.95	12.29
Prenatal Care Visits	11.88	3.56
Month Prenatal Care Began	2.32	1.27
Birth Order	2.38	1.44
Weather		
Average Temperature	12.98	5.04
Average Rainfall	2.90	1.01
Average Snowfall	1.81	2.20
Average Sunlight	15803.66	2065.49
Average Days in Temperature Bins		
(-Inf,-8]	6.11	10.08
(-8,-4]	9.50	10.15
(-4,0]	19.28	15.86
(0,24]	195.43	40.96
(24,28]	32.64	29.01
(28,32]	12.12	21.79
(32, Inf]	1.29	7.70
Average Days in Rainfall (in mm) Bins		
[0,10]	251.39	18.36
(10,20]	16.02	6.82
(20,50]	8.23	4.80
(50,100]	0.70	1.02
(100, Inf]	0.03	0.19
Number of Observations		19,865,677

Source: NCHS, NLDAS, NCEI (1989-2004)

Table 2: Summary Statistics for Birth Data and Heat Shocks

Variables	No Shock Mean	SD	Heat Shock Mean	SD	Normalized Difference
Infant Characteristics					
Birth Weight in grams	3403.15	483.16	3390.19	477.19	-0.02
Standardized Birth Weight	-0.04	1.02	-0.06	1.01	-0.01
Low Birth Weight	0.04	0.19	0.04	0.19	-0.00
Small for Gestational Age	0.11	0.31	0.11	0.31	0.01
5-min. APGAR Score	8.97	0.55	8.96	0.56	-0.02
Gestation Length	39.12	1.74	39.09	1.75	-0.01
Preterm Birth	0.15	0.35	0.15	0.36	0.01
Assisted Ventilation	0.02	0.13	0.02	0.13	-0.01
Male	0.51	0.50	0.51	0.50	-0.00
Mothers Characteristics					
Mother's Age	28.29	5.73	27.51	5.73	-0.10
Married	0.81	0.39	0.80	0.40	-0.01
Race - White	0.82	0.38	0.81	0.40	-0.03
Race - Black	0.13	0.34	0.15	0.36	0.05
Race - Other	0.05	0.21	0.04	0.20	-0.02
Non-Hispanic	0.89	0.32	0.84	0.36	-0.09
Hispanic - Mexican	0.05	0.22	0.10	0.30	0.14
Hispanic - Other	0.06	0.24	0.05	0.23	-0.03
Education - <HS	0.03	0.18	0.04	0.19	0.01
Education - Highschool	0.09	0.29	0.11	0.32	0.04
Education - some College	0.55	0.50	0.57	0.49	0.03
Education - College +	0.32	0.47	0.28	0.45	-0.07
Smoking during Pregnancy	0.11	0.31	0.10	0.30	-0.03
Weight Gain in pounds	30.92	12.24	31.28	12.66	0.02
Prenatal Care Visits	11.85	3.55	12.09	3.72	0.05
Month Prenatal Care Began	2.33	1.27	2.27	1.32	-0.03
Birth Order	2.38	1.44	2.36	1.42	-0.01
Weather					
Average Temperature	12.36	4.70	18.64	4.47	0.97
Average Rainfall	2.95	0.94	2.43	1.38	-0.31
Average Snowfall	1.95	2.25	0.48	0.97	-0.60
Average Sunlight	15560.66	1892.58	18019.19	2249.17	0.84
Number of Observations		17,902,118		1,963,559	

Source: NCHS, NLDAS, NCEI (1989-2004)

Over the 25 years that we observe, there is a slight increase in average temperatures, see figure 1. Particularly average temperatures in the winter months seem to have increased. But also summer averages show a subtle increase. To motivate our shock definition, we are first interested in the effects of hot and cold temperatures simultaneously. We use the mean daily temperature and divide it into the following segments: $< 8°C, (-8°C, -4°C], (-4°C, 0°C], (0°C, 24°C], (24°C, 28°C], (28°C, 32°C], > 32°C$. For the analysis, we set $(0°C, 24°C]$ as the reference temperature bin, since most days lie within these boundaries (see figure A.3). For each mother, we calculate the number of days in each exposure bin during pregnancy. To further examine sensitivity by the timing of the shock, we calculate the number of days of exposure in each trimester of the pregnancy.

In the main part of the analysis, we want to analyze the effect of exposure to extreme temperature events. Since especially long-term exposure to heat or extended heat periods are found to have negative effects, we define a heat shock as the average temperature of 5 consecutive days exceeding the 0.85 percentile of historic July and August temperatures and at least 30 degrees Celsius. Specifically, we assess whether the 5-day average temperature exceeds that of the 9th hottest day recorded in each county between 1979 and 1988. In addition, we require that the temperature surpasses a threshold of at least 30 degrees, which serves to exclude counties that do not typically experience extremely hot days.

Table 2 shows that mothers who experience a heat shock during pregnancy are on average very similar to those who do not experience such a shock. The normalized difference indicates no substantial difference in any of the infant characteristics or most of the maternal characteristics. There is slightly higher education for those mother's not experiencing a shock. Black mothers are more likely to experience a heat shock, the same holds for mothers of Mexican origin. There are substantial differences regarding average weather. The average temperature in pregnancies by mothers experiencing a shock is nearly 8 $°C$ higher, whereas average rainfall is substantially lower than for mothers not experiencing a shock. In particular, most mothers who are pregnant during summer are exposed to heat shocks.

The second weather condition we consider is daily rainfall (in mm). Similar to temperature, we first investigate the non-linear effects of rainfall. We count the number of days per pregnancy or trimester that each mother is exposed to the following rainfall bins: $0 - 10$mm, $10 - 20$mm, $20 - 50$mm, $50 - 100$mm, > 100mm. These reflect a range from very light rain to very heavy rainfall. Overall, average rainfall does not change during the 25 years of interest and shows expected seasonality. As expected, most days lie in the low precipitation bin of $0 - 10mm$, which serves as a reference category for our analysis. As figure 2 shows, there is quite some deviation from the mean average rainfall. Other than for temperature, we do not see strong seasonality. While the average daily rainfall mostly lies below 5 mm, there are large outliers. Figure A.4 shows the average distribution of days in a certain rainfall bin. On average, counties experience less than a day in the most extreme rainfall bin (more than 100mm per day), indicating that this is a very rare event. Usually, counties mostly experience days of no or very light rainfall.

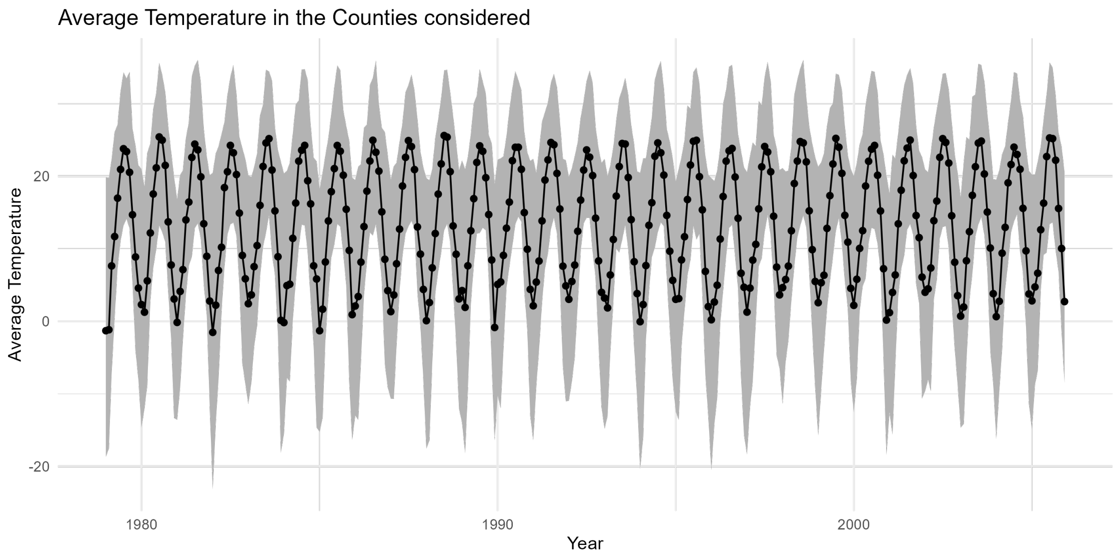

Figure 1: Average Temperature

Note: This plot shows the average daily temperature per month in the counties considered. The grey area indicates the range for the maximum and minimum value for average temperature in a certain month.

3 Empirical Strategy

3.1 Setup and Notation

For the analysis of heterogeneity, we focus on extreme weather shocks. We want to estimate the causal effects of these binary weather shocks in the potential outcomes framework by Imbens and Rubin (2015). For each observation i we define the outcome under potential treatment as $Y_i^{(T_i)}$, $T_i \in \{0,1\}$, where $Y_i^{(0)}$ denotes the potential outcome if individual i did not receive the treatment and $Y_i^{(1)}$ denotes the potential outcome if i did receive the treatment. In any case, we can only observe the realized outcome $Y_i^{obs} = T_i Y_i^{(1)} + (1 - T_i) Y_i^{(0)}$, since both potential outcomes can never be observed together. The propensity score $p(X_i) = P(T_i = 1|X_i)$ denotes the probability of receiving the treatment, giving the observable covariates of individual i. The average treatment effect (ATE) is defined as $\tau_{ATE} = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)}]$, which is the expected difference between the potential outcomes. The main interest in this analysis are treatment effects that differ across individuals by their observable characteristics. Therefore, we focus on the conditional average treatment effect (CATE), defined as

$$\tau(x) = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)}|X_i = x]. \tag{1}$$

Given that the heat shock is not a random shock, due to its correlation with the mothers' location of residence and the minimum required temperature of 30 °C, we assume unconfoundedness $\{Y_i^{(0)}, Y_i^{(1)}\} \perp T_i|X_i$ and overlap $0 < P(T_i = 1|X_i) < 1$ in the covariate distributions to be able to identify the ATE and CATE.

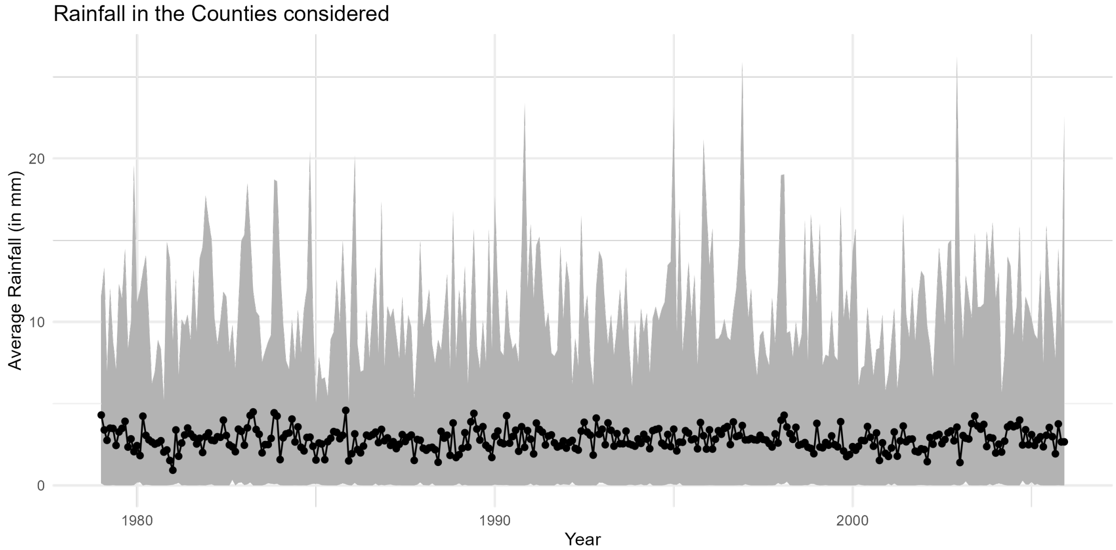

Figure 2: Average Rainfall (in mm)

Note: This plot shows the average daily rainfall per month in the counties considered. The grey area indicates the range for the maximum and minimum value for average rainfall in a certain month.

3.2 Identification Challenges

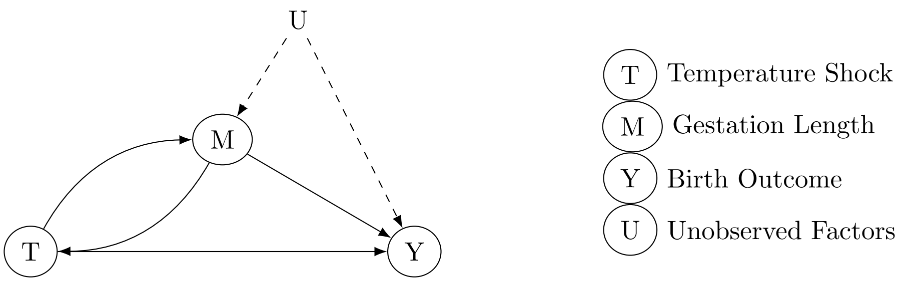

Figure 3: DAG representation of the relation between Outcomes, Gestation and Temperature Exposure

A problem we face in the identification of causal effects of shocks during pregnancy on health at birth is the mechanical correlation between gestation length and our temperature exposure measures. For a longer gestation length, the probability of exposure to a temperature shock increases. Figure 3 is a simplified representation of the problem at hand. Temperature exposure has a direct effect on the birth outcome and an indirect effect via gestation length. There is a simultaneous interaction between temperature exposure and gestation length. Gestation acts as a mediator on the path $T \rightarrow M \rightarrow Y$ and as a confounder on the path $T \leftarrow M \rightarrow Y$. In case of gestation length only acting as a confounder affecting both the treatment and the outcome, one would want to control for it. However, given that gestation is also a possible outcome of the treatment it is a bad control. Additionally, there might be other unobserved factors that are common causes of gestation length and birth outcomes. We denote this unobserved factor as U in figure 3. U could for example be genetic factors or birth defects. In the literature, M is referred to as a collider (Imbens, 2020). Controlling for gestation length would induce some sort of selection bias, referred to as collider stratification bias. When conditioning on gestation length, we open the causal path from $T \rightarrow M \leftarrow U \rightarrow Y$, which is a source of bias in the estimates (Imbens, 2020)[6]. So in the given setup, both conditioning and not conditioning on gestation length induces bias.

The literature usually attempts to solve this problem by assuming fixed pregnancy length, i.e., 40 weeks or 9 months of pregnancy (Bratti et al., 2021, Chen et al., 2020, Deschênes et al., 2009, Le and Nguyen, 2021). Only Currie and Rossin-Slater (2013) discuss the problem of the mechanical correlation between exposure and gestation length. Assuming fixed pregnancy length imposes bias, depending on how the gestation period is measured. As Currie and Rossin-Slater (2013) point out, counting backward from the date of birth shifts trimester borders, where exposure in the third trimester is most likely overstated and the strategy measures exposure that happened before the pregnancy even started. While the second strategy does not shift trimester borders, it measures exposure to extreme events after pregnancy, therefore after measurement of the outcomes of interest. Some individuals with shorter gestation lengths are therefore part of the treatment group, even though they should be in the control group, given that exposure happened after birth. This will lead to an overestimation of the treatment effect. To handle the problem of this mechanical correlation, we confine outcomes to birth weight standardized for gestational age, SGA, and 5-minute Apgar score. This way, the outcome itself attempts to control for gestational age. Appendix B illustrates the problem of the mechanical correlation between gestation length and temperature exposure.

3.3 Treatment Effect Estimation using Causal Forests

To estimate heterogeneous treatment effects of the binary shocks, we use the causal forest by Athey et al. (2019). The causal forest builds on a partially linear model by Robinson (1988) with constant treatment effect $\tau(x) = \tau$ for all $x \in \mathcal{X}$:

$$Y_i = g(X_i) + T_i\tau + \epsilon_i. \tag{2}$$

Let $m(X_i) = E[Y_i|X_i]$ be the expected outcome and $p(X_i)$ the propensity score as defined before, then the model can be rewritten as

$$Y_i - m(X_i) = (T_i - p(X_i))\tau + \epsilon_i \tag{3}$$

and one can estimate τ using the following estimator, which is a semiparametrically efficient estimator under unconfoundedness (Robinson, 1988, Chernozhukov et al., 2018b)

$$\hat{\tau} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}(X_i)\right)\left(T_i - \hat{p}(X_i)\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(T_i - \hat{p}(X_i)\right)^2}. \tag{4}$$

[6] A prominent example for collider bias is the low birth weight paradox. Several studies find that infants born to smokers have higher risks for LBW and infant mortality than infants born to non-smokers. However, among infants born at LBW, mortality is lower for those babies born to smokers, which would imply some beneficial effects of smoking (Hernández-Díaz et al., 2006). Similar to gestation length in figure 3, LBW is a collider. Smoking is not beneficial, but LBW infants born to smokers have lower mortality than LBW infants born to smokers since LBW of babies born to non-smokers is due to more detrimental causes associated with higher infant mortality, like birth defects.

Athey et al. (2019) build on this idea, but allow the treatment effect to differ with observables. To estimate $\tau(x)$ they rely on equation (4), but assume the treatment effect to be constant in a sufficiently small neighborhood $N(x)$. They use the random forest (Breiman, 2001) to find these neighborhoods where the treatment effect is assumed to be constant.

Random forests are a combination of many full-grown regression trees (Breiman, 2001). Regression trees split the given feature space into non-overlapping rectangular regions, also referred to as leaves. Each tree $b = 1, \ldots, n$ is grown on a random subset sampled from the original feature space $\mathcal{S}_b \subseteq 1, \ldots, n$, by greedy recursive partitioning, where we recursively choose splits, which minimize the prediction error in the sample. For every observation that falls into a certain leaf L, a tree returns the mean of all response values of training observations that fall into the same leaf. Averaging over all those trees' predictions leads to the prediction of the random forest for $\mu(x) = \mathrm{E}[Y_i|X_i = x]$:

$$\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{Y_i 1(\{X_i \in L_b(x), i \in \mathcal{S}_b\})}{|\{X_i \in L_b(x), i \in \mathcal{S}_b\}|}, \tag{5}$$

where L_b is the leaf containing x in tree b.

Thinking of the random forest as an adaptive kernel method, with data-adaptive kernel $\alpha_i(x)$, we can equivalently write

$$\hat{\mu}(x) = \sum_{i=1}^{n} \alpha_i(x) Y_i, \quad \alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{1(\{X_i \in L_b(x), i \in \mathcal{S}_b\})}{|\{X_i \in L_b(x), i \in \mathcal{S}_b\}|}, \tag{6}$$

where the weights $\alpha_i(x)$ measures how often the i-th observation falls into the same leaf as x. These data-adaptive kernel weights can be used to characterize the neighborhood $N(x)$, and estimate the CATE function as follows:

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x) \left(Y_i - \hat{m}(X_i)\right) \left(T_i - \hat{p}(X_i)\right)}{\sum_{i=1}^{n} \alpha_i(x) \left(T_i - \hat{p}(X_i)\right)^2}. \tag{7}$$

In practice, the `grf` implementation of the causal forest fits two separate regression forests to estimate unknown propensity scores $\hat{p}(x)$ and marginal outcomes $\hat{m}(x)$ on and makes out-of-bag predictions. With these estimates it residualizes outcomes and treatments to grow causal trees. The splits of the trees are chosen, so that the treatment effect within each leaf is homogeneous, while there is heterogeneity in treatment effects between leaves. For details see Athey et al. (2019). Several trees are grown on many subsamples of the data and averaged to build a causal forest, which is then used to derive the weights for each observation $\alpha_i(x)$.

The average treatment effect (ATE) can be estimated by a variant of doubly robust augmented inverse-propensity weighting (AIPW):

$$\hat{\tau}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}(X_i) + \frac{T_i - \hat{p}(X_i)}{\hat{p}(X_i)(1-\hat{p}(X_i))}\left(Y_i - \hat{m}(X_i) - (T - \hat{p}(X_i))\hat{\tau}(X_i)\right)\right) \tag{8}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + \frac{T_i}{p(\hat{X}_i)}(Y_i - \hat{\mu}_1(X_i)) - \frac{1-T_i}{1-p(\hat{X}_i)}(Y_i - \hat{\mu}_0(X_i))\right), \tag{9}$$

where $\mu_{(T)}(X_i) = E[Y_i|X_i, T_i] = m(X_i) + (T_i - p(X_i))\tau(X_i)$. Cross-fitting is used to avoid overfitting, for both the ATE and CATE prediction. The propensity score and outcomes models are trained on one part of the data (auxiliary sample), whereas the causal forest is estimated on another part of the data (main sample). To not loose half of the observations, we can switch the roles of the main and the auxiliary samples and report averaged results. See Chernozhukov et al. (2018a) for details on cross-fitting.

3.4 Assessing Treatment Effect Heterogeneity

To judge whether there is heterogeneity and whether the estimated $\tau(x)$ on average captures the ATE well, we use a calibration test. For this test we run the following regression, inspired by the BLP in Chernozhukov et al. (2018b), but adapted for estimation in observational settings:

$$Y - \hat{m}(x) = \beta_1(T_i - \hat{p}(X_i))\bar{\hat{\tau}}(x) + \beta_2(T_i - \hat{p}(X_i)) * (\hat{\tau}(x) - \bar{\hat{\tau}}(x)), \tag{10}$$

where $(T_i - \hat{p}(X_i))\bar{\hat{\tau}}(x)$ refers to the mean forest prediction and $(T_i - \hat{p}(X_i)) * (\hat{\tau}(x) - \bar{\hat{\tau}}(x))$ to the differential forest prediction. A coefficient β_1 close to 1 suggests that the mean forest prediction is correct. The p-value of the β_2 coefficient acts as an omnibus test for the presence of heterogeneity. If β_2 is significantly different from 0, we can reject the null of no heterogeneity.

Following Zenzes (2022), we will additionally use an effect decomposition to uncover differences in treatment effects solely attributable to a single variable. The decomposition allows us to detect mothers' characteristics for which we can find large differences in treatment effects while keeping all other characteristics fixed. This decomposition is based on a decomposition approach by Chernozhukov et al. (2013) that makes use of the counterfactual distribution. We split the sample into non-overlapping populations based on a variable of interest and decompose the treatment effect for individuals in these populations. We decompose the treatment effect difference into the structural effect, which arises due to differences in treatment effect for individuals with the same characteristics, and the compositional effect arising from differences in the characteristics between the groups. Using this decomposition on the estimated heterogeneous treatment effect allows for detecting the direct effect of a variable on the treatment effect (structural effect) and helps to identify sources of heterogeneity.

Considering two non-overlapping populations $j, k \in \mathcal{K}$, the observed differences in the treatment effects can be decomposed as

$$\underbrace{F_{\tau\langle k|k\rangle}(y) - F_{\tau\langle j|j\rangle}(y)}_{\textit{Total Effect}} = \underbrace{[F_{\tau\langle k|k\rangle}(y) - F_{\tau\langle j|k\rangle}(y)]}_{\textit{Structural Effect}} + \underbrace{[F_{\tau\langle j|k\rangle}(y) - F_{\tau\langle j|j\rangle}(y)]}_{\textit{Compositional Effect}}. \tag{11}$$

$F_{\tau\langle k|k\rangle}(y)$ is the observed distribution function of $\tau(x)$ for population k. $F_{\tau\langle j|k\rangle}(y)$ represents the counterfactual distribution function of τ that would have prevailed for population k, if they faced populations j's distribution of τ. It can also be interpreted as the distribution function of τ that would have prevailed for population j, if they faced population k's covariate distribution. The derived decomposition shows the variation along the whole distribution of treatment effects. For details on the procedure, see algorithm 1 in appendix C and Zenzes (2022).

3.5 Fixed Effects Regression Analysis

To be able to compare estimates with previous studies and to get a better understanding of how different temperature bins affect health at birth to determine meaningful binary weather shocks, we model the effect of temperature and rain exposure during pregnancy as follows:

$$Y_{icmy} = \alpha + \sum_{k=1}^{n} \beta_k T_{icmy} + \gamma W_{cmy} + \theta X_{icmy} + \eta_c + \mu_m + \nu_y + \sigma_s + \mu_m x \eta_c + \epsilon_{icmy}, \tag{12}$$

where the dependent variable Y_{icmy} is the birth outcome of infant i, in county c and month m and year y. The key variable of interest is T_{icmy}, which is the count of days during pregnancy, where the temperature or rainfall lies in bin k. The vector W_{cmy} contains several weather controls, such as average rainfall, snowfall, and sunshine in county c, month m, and year y. X_{icmy} contains a set of maternal and child characteristics, such as mother's age, educational attainment, marital status, sex of child, prenatal care, and birth order. Additionally, we control for several fixed effects, on the county of residence η_c, month of birth μ_m, year of birth ν_y, and state level σ_s. Standard errors are clustered at the county level. By controlling for these fixed effects, we control for seasonality in birth outcomes, structural differences between places of residence and the selection into places of residence and hence climate of residence.

To further assess how different timing of certain shocks during pregnancy alter the effect, we estimate the following specification:

$$Y_{icmy} = \alpha + \sum_{k=1}^{n} \beta_k^{TR} T_{icmy}^{TR} + \gamma W_{cmy} + \theta X_{icmy} + \eta_c + \mu_m + \nu_y + \sigma_s + \mu_m x \eta_c + \epsilon_{icmy}, \tag{13}$$

where we now look at the effects in the different trimesters of pregnancy. The variables of interest T_{icmy}^{TR} either corresponds to T_{icmy}^{TR1}, T_{icmy}^{TR2}, or T_{icmy}^{TR3} which are counting the days the temperature (or rainfall) falls in bin k in each trimester of pregnancy. $TR1$, $TR2$, and $TR3$ are the indicators for the first, second, and third trimesters of pregnancy respectively. All other variables are defined as in equation (12).

4 Results

4.1 Temperature Effects

Table 3 presents estimates of the effect of temperature on several birth outcomes of interest. The results displayed follow equation (12), so the estimates show the effect of an additional day in a certain temperature bin compared to a day in the reference bin, which is 0-24°C. Column (1) shows the effect of temperature on standardized birth weight. Consistent with the literature, we find significant, but economically small negative effects of an additional day in hot temperatures on standardized birth weight, compared to a day in the reference temperature. Especially very hot temperatures show a significant reduction of -0.001 (corresponding to -0.48 grams), whereas the effect of exposure to less hot temperatures is smaller. On the other extreme of the temperature distribution, we observe positive effects, but these do not increase with lower temperatures. Controlling for mother's characteristics (column (2)) and average weather in the pregnancy months does not change the estimates much. Regarding the effect on SGA birth (columns (3) and (4)), we observe effects very close to zero. However, the same pattern as for birth weight is visible. Hot temperatures increase SGA births, whereas lower temperatures decrease SGA births. For the standardized Apgar score, we see mostly insignificant and very small effects of exposure to temperature extremes. Only extreme heat (above 32°C) shows significant positive effects on the Apgar score in both specifications.

In summary, there are negative effects of exposure to hot temperatures and small positive effects of cold exposure on birth weight. Effects are particularly strong for extreme temperatures above 32 °C. Overall, effects most profound in the second and third trimesters (see appendix D). This is well in line with findings from the literature. Comparing the effect size found to other studies, we find very similar economically small effects as previous studies. Deschênes et al. (2009) also study births in the US and find reductions in birth weight of around 0.003 to 0.009 percent for exposure to temperatures higher than 30 °C. This corresponds to changes of 0.102 to 0.3 grams. Chen et al. (2020) study births in rural China and find reductions of birth weight by 1.66 grams for temperatures above 28 °C. Estimates from both studies perfectly align with our average effects found. We see slightly different estimates, given that our temperature bins are not exactly comparable.

4.2 Heat Shock Effects

We want to understand possible heterogeneity in the effect of temperature on birth outcomes. This analysis focuses on temperature exposure only, given that rainfall exposure has no significant effect on the outcomes of interest (see appendix E). The results in table 3 motivate our shock definition. Since exposure to hot temperatures showed the largest, robust effects on birth weight, we will focus on the heterogeneity in the effect of extreme temperature shocks on birth weight-related outcomes.

When estimating the propensity score for the entire sample we encounter problems of weak overlap. A lot of mass is concentrated at zero. This is because the temperature shock is not exogenous, given its strong

Table 3: Temperature Effects

Dependent Variables:	Standardized Birth Weight		SGA		Standardized Apgar Score	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Avg. Temp. <-8°C	0.4083***	0.1418	-0.0659***	-0.0027	-0.2954	-0.1987
	(0.0840)	(0.1174)	(0.0243)	(0.0252)	(0.2112)	(0.2436)
Avg. Temp. -8 - -4 °C	0.5469***	0.3567***	-0.1315***	-0.0970***	0.1979	0.3773
	(0.0973)	(0.0890)	(0.0238)	(0.0249)	(0.2700)	(0.2736)
Avg. Temp. -4 - 0 °C	0.6056***	0.2629***	-0.1264***	-0.0499***	-0.0059	0.1224
	(0.0703)	(0.0968)	(0.0156)	(0.0192)	(0.1449)	(0.1592)
Avg. Temp. 24 - 28 °C	-0.1439**	-0.3211***	-0.0127	0.0357**	-0.1129*	-0.1012
	(0.0644)	(0.0787)	(0.0123)	(0.0149)	(0.0625)	(0.0708)
Avg. Temp. 28 - 32 °C	-0.2833***	-0.4122***	0.0123	0.0436**	-0.0714	-0.0196
	(0.0823)	(0.0932)	(0.0163)	(0.0177)	(0.0914)	(0.1175)
Avg. Temp. >32 °C	-1.0404***	-1.1810***	0.1349***	0.1758***	0.4340**	0.5180***
	(0.1109)	(0.1142)	(0.0263)	(0.0217)	(0.1734)	(0.1806)
Weather Controls		Yes		Yes		Yes
Mother's Characteristics		Yes		Yes		Yes
Fixed-effects						
Year of Birth	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth	Yes	Yes	Yes	Yes	Yes	Yes
County	Yes	Yes	Yes	Yes	Yes	Yes
State	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth-County	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	19,865,677	18,595,704	19,865,677	18,595,704	19,865,677	18,595,704
R^2	0.0092	0.0900	0.0039	0.0376	0.0265	0.0283
Within R^2	0.0000	0.0815	0.0000	0.0338	0.0000	0.0025

Notes: Entries show the coefficient on the relevant Temp. exposure measure, scaled by 1000 for ease of reading. This means that a one-unit change in *Avg. Temperature <-8° C* is associated with a 0.0004 unit change in standardized birth weight in column (1), holding all other variables constant. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension.
Standard errors clustered by County of residence in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

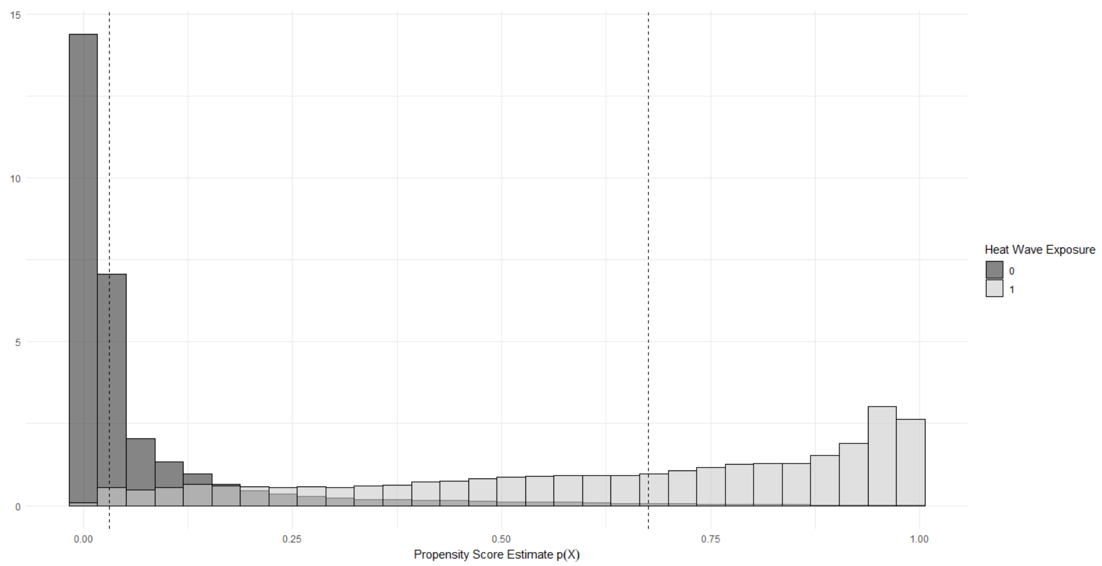

Figure 4: Propensity Score estimate for heat shock exposure

Note: The propensity of heat shock exposure was estimated using a random forest. Dashed lines indicate the borders to trim the propensity score for sufficient overlap.

correlation with the location and some locations never experience a heat shock. To have better overlap, we therefore exclude states, that either show no exposure at all or very weak overlap. The excluded states are Connecticut, Delaware, Maine, Massachusetts, New Hampshire, New York, Rhode Island, Pennsylvania, Vermont, and Virginia. So, we exclude all states in the North East. To further ensure sufficient overlap, we employ an asymmetric propensity score trimming approach by Stürmer et al. (2010). They first discard all observations in regions without overlap in estimated propensities of treated and untreated observations. In the second step, they derive an upper and lower bound to trim the propensity score estimates. For the lower end, they use the lowest 0.01 percentile of the propensity score estimated for treated observations. At the upper end of the propensity score distribution, they use a restriction based on the highest 0.99 percentile of propensity scores in the untreated population. Following this approach results in the borders indicated in figure 4.

Table 4: Average Treatment Effect of Heat Shock during pregnancy

	ATE	Std. Error
Standardized Birth Weight	-0.0124***	0.0042
SGA	0.0042***	0.0014

Notes: ATE is estimated using the inverse-propensity weighting in equation (8).
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Figure 5 shows a histogram of the estimated CATE of the heat shock on standardized birth weight and the ATE with corresponding 95% confidence intervals. The average treatment effect is -0.0124 (see table 4) and is significantly different from zero. It corresponds to a reduction of around 6 grams of birth weight. The calibration test in table 5 indicates heterogeneity being present. The coefficient for 'mean.forest.prediction' is close to 1, suggesting, that the mean forest prediction is correct. We can additionally reject the null of no

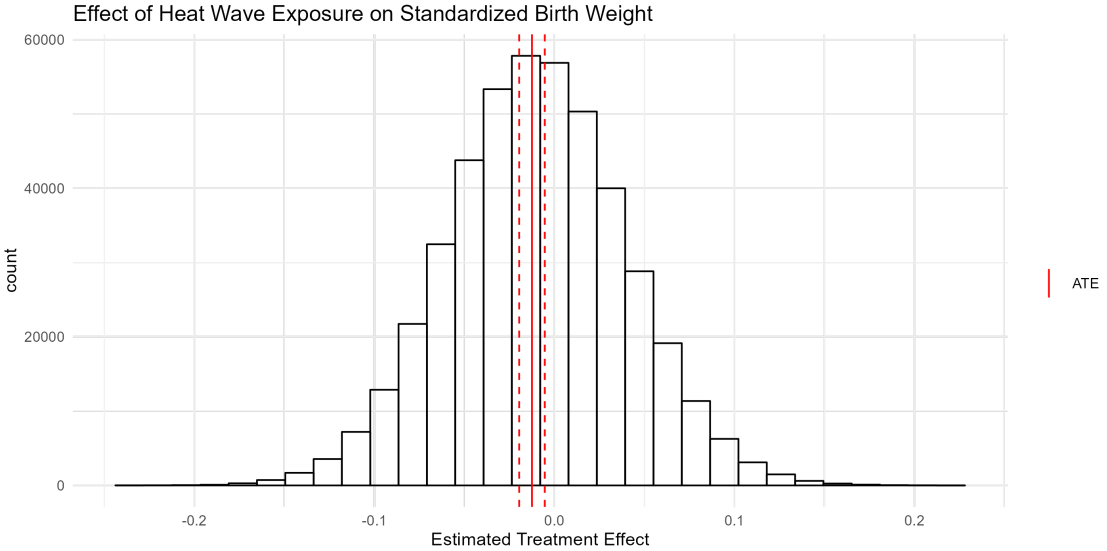

Figure 5: CATE of heat shock on standardized birth weight

Note: The histogram shows the estimated CATE of heat shock on standardized birth weight. CATE is estimated using a causal forest using the `grf` R-package. The solid red line shows the ATE and corresponding 95% confidence intervals. These are estimated using the `grf` R-package by doubly robust augmented inverse propensity weighting.

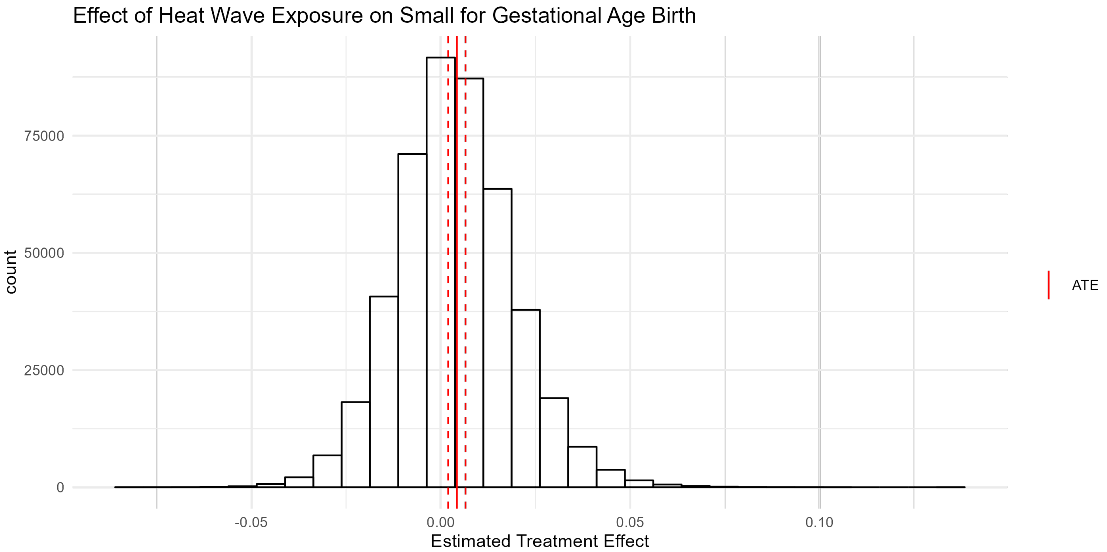

Figure 6: CATE of heat shock on SGA birth

Note: The histogram shows the estimated CATE of heat shock on SGA birth. CATE is estimated using a causal forest using the `grf` R-package. The solid red line shows the ATE and corresponding 95% confidence intervals. These are estimated using the `grf` R-package by doubly robust augmented inverse propensity weighting.

heterogeneity being present since the coefficient of 'differential.forest.prediction' is significantly greater than 0.

Figure 6 displays the CATE estimates for the effect of a heat shock in small for gestational age birth. We observe a small but significant increase in SGA rate of 0.0042 (see table 4). So, on average, the heat shock affects infants on the lower end of the birth weight distribution, which are most vulnerable. The results of the calibration test in table 5 show that the average treatment effect estimate is correct, however the test cannot detect substantial heterogeneity in the effect.

Table 5: Calibration Test Results

Birth Outcome	Variables	Estimate	Std. Error	t value	Pr(>t)
Std. Birth Weight	mean.forest.prediction	0.9520	0.4352	2.1877	0.0143
	differential.forest.prediction	0.2056	0.0871	2.3598	0.0091
SGA Birth	mean.forest.prediction	0.9653	0.3581	2.6953	0.0035
	differential.forest.prediction	-0.0965	0.0974	-0.9906	0.8391

Notes: Best linear fit using forest predictions (on held-out data) as well as the mean forest prediction as regressors, along with standard errors, as described in equation (10).

Given that the literature agrees on negative effects of in-utero exposure to heat events on health at birth, the results displayed in figure 5 are questionable. While it makes sense, that some mothers can cope better with heat wave exposure than others, questions arise from the positive effect found. These results were robust to different propensity score estimation and trimming approaches and inclusion and exclusion of several characteristics of the mother. While we do not support the idea of positive effects of heat waves exposure, we still think it is appropriate to analyze differences in groups with strong negative (most affected) and strong positive (least affected) groups to uncover structural differences between them.

To be able to characterize the group most vulnerable to heat shocks, we will compare the average characteristics of 10% most and 10% least affected groups. Most affected refers to mothers with the strongest negative effect and least affected to mothers with estimated positive effect. Table 6 shows this comparison. Most and least affected groups differ a lot by their characteristics, especially when it comes to mother's race, Hispanic origin, and education. The 10% most affected group comprises a lot more black and Mexican mothers than the least affected group. Also, the education level of the most affected group is lower than that of the least affected group. On average, most affected mothers are 2 years younger and less likely to be married. We don't observe a large difference in birth order or smoking during pregnancy, but lower prenatal visits and lower weight gain in the group of most affected mothers. Comparing average exposure to different weather measures, we do not find a large difference between the two groups.

To further evaluate drivers of heterogeneity, we decompose differences in estimated treatment effects for groups into the effect of a single variable, while keeping the other characteristics comparable. So the decomposition follows the procedure in 1. Given that the differences found between most and least affected groups (table 6) were mostly regarding the mother's age, race, Hispanic origin, and weight gain, we restrict the decomposition analysis to these factors. Figure 7 shows the full decomposition for mother's age. For all

Table 6: 10% most and least affected mothers

Variables	10% most affected Mean	SD	10% least affected Mean	SD	Normalized Difference
Mothers Characteristics					
Mother's Age	26.61	6.03	28.61	6.43	-0.23
Married	0.73	0.45	0.85	0.36	-0.22
Race - White	0.69	0.46	0.85	0.35	-0.28
Race - Black	0.28	0.45	0.11	0.32	0.30
Race - Other	0.03	0.17	0.03	0.18	-0.01
Non-Hispanic	0.81	0.39	0.90	0.30	-0.19
Hispanic - Mexican	0.15	0.36	0.03	0.17	0.30
Hispanic - Other	0.04	0.20	0.07	0.25	-0.09
Education - <HS	0.07	0.25	0.02	0.14	0.16
Education - Highschool	0.15	0.36	0.08	0.26	0.17
Education - some College	0.53	0.50	0.56	0.50	-0.04
Education - College +	0.25	0.43	0.35	0.48	-0.15
Smoking during Pregnancy	0.07	0.26	0.10	0.29	-0.05
Cigarettes during Pregnancy	1.03	6.02	1.47	7.38	-0.05
Weight Gain in pounds	29.01	12.06	32.89	14.33	-0.21
Prenatal Care Visits	11.45	3.77	12.79	3.90	-0.25
Month Prenatal Care Began	2.51	1.53	2.16	1.21	0.18
Birth Order	2.30	1.34	2.29	1.44	0.00
Infant Characteristics					
Birth Weight in grams	3339.78	480.77	3414.37	480.37	-0.11
Standardized Birth Weight	-0.14	1.01	-0.01	1.01	-0.09
Low Birth Weight	0.05	0.21	0.04	0.19	0.03
Small for Gestational Age	0.13	0.33	0.10	0.30	0.05
5-min. APGAR Score	8.94	0.55	8.96	0.54	-0.02
Gestation Length	38.95	1.81	39.10	1.75	-0.06
Preterm Birth	0.17	0.38	0.15	0.36	0.05
Assisted Ventilation	0.02	0.14	0.02	0.13	0.00
Male	0.51	0.50	0.51	0.50	-0.00
Weather					
Average Temperature	16.46	4.78	17.24	4.75	-0.12
Average Rainfall	2.82	1.22	3.02	1.13	-0.12
Average Snowfall	0.93	1.92	0.72	1.50	0.09
Average Sunlight	16902.66	1997.23	17159.06	1894.14	-0.09

other possibly modifying factors, we will only display the structural effect, as it is the main effect of interest.

Figure 7 follows the implementation steps as described in procedure 1, repeated 5 times using different random splits into a main and auxiliary sample. The differences refer to the baseline group of mothers aged 10-23. We report confidence bands for $\alpha = 0.05$, meaning we use bootstrap confidence bands for confidence level 0.975 but discount for splitting uncertainty, resulting in a confidence level of 0.95. The x-axis of the plot shows the quantile index, lower quantiles show the most harmful effects. The y-axis displays the differences in treatment effects. The top left plot shows the quantiles of the observed CATE cumulative distribution (solid line) and counterfactual distribution that would have prevailed for other groups if they faced the reference group's distribution of characteristics (dashed line). The top right shows the total effect, which is the observed difference between each group's CATE distribution and the CATE distribution of the baseline group. The bottom left displays the structural effect, which is our main effect of interest. It shows the difference only associated with the mother's age. Characteristics of groups are kept comparable by using the counterfactual distribution. This difference is always relative to the reference group. The bottom right shows the compositional effect, which is the difference associated with differences in group characteristics, leaving out mother's age.

We find a weak modifying effect for mother's age in the effect of a heat shock on standardized birth weight. The total effect indicates significant differences between mothers younger than 27 and older than 27. Similarly, the compositional effect shows that there are strong differences in treatment effects given differences in group composition, especially for older mothers. However, the structural effect does not reveal significant differences between mothers in different age groups. For mothers aged 27 to 32, we can detect weakly significant mitigating effects associated with age. At lower quantiles, there are significant amplifying effects for older mothers, but these fade with increasing quantile index. For mothers aged 23 to 27, we cannot find any difference compared to our reference group.

While we detected significant differences between most and least affected groups regarding their race, Hispanic origin, and education, the decomposition does not confirm this observation. We do observe that differences in race tend to amplify effects, but these are not significantly different from zero (figure 8). Similarly, we cannot detect significant modification by Mexican or other Hispanic origins (figure 9) or any level of mother's education (figure 10). Only the decomposition for weight gain during pregnancy reveals weak mitigating effects for excessive weight gain (figure 11).

In summary, this means that while we can detect differences in most and least affected groups, we are unable to clearly describe the drivers of heterogeneity found, apart from a weak effect modification by age and weight gain. This means that heterogeneity is not driven by race, Hispanic origin, education, or corresponding associated factors on their own. Keeping other characteristics comparable fully removes any modification effect attributable to race or other strongly associated factors. The actual drivers might be a combination of certain factors or circumstances that are more complex to measure. The decomposition thus shows that we can rule out the mother's race, Hispanic origin, and education as single driving factors of heterogeneity.

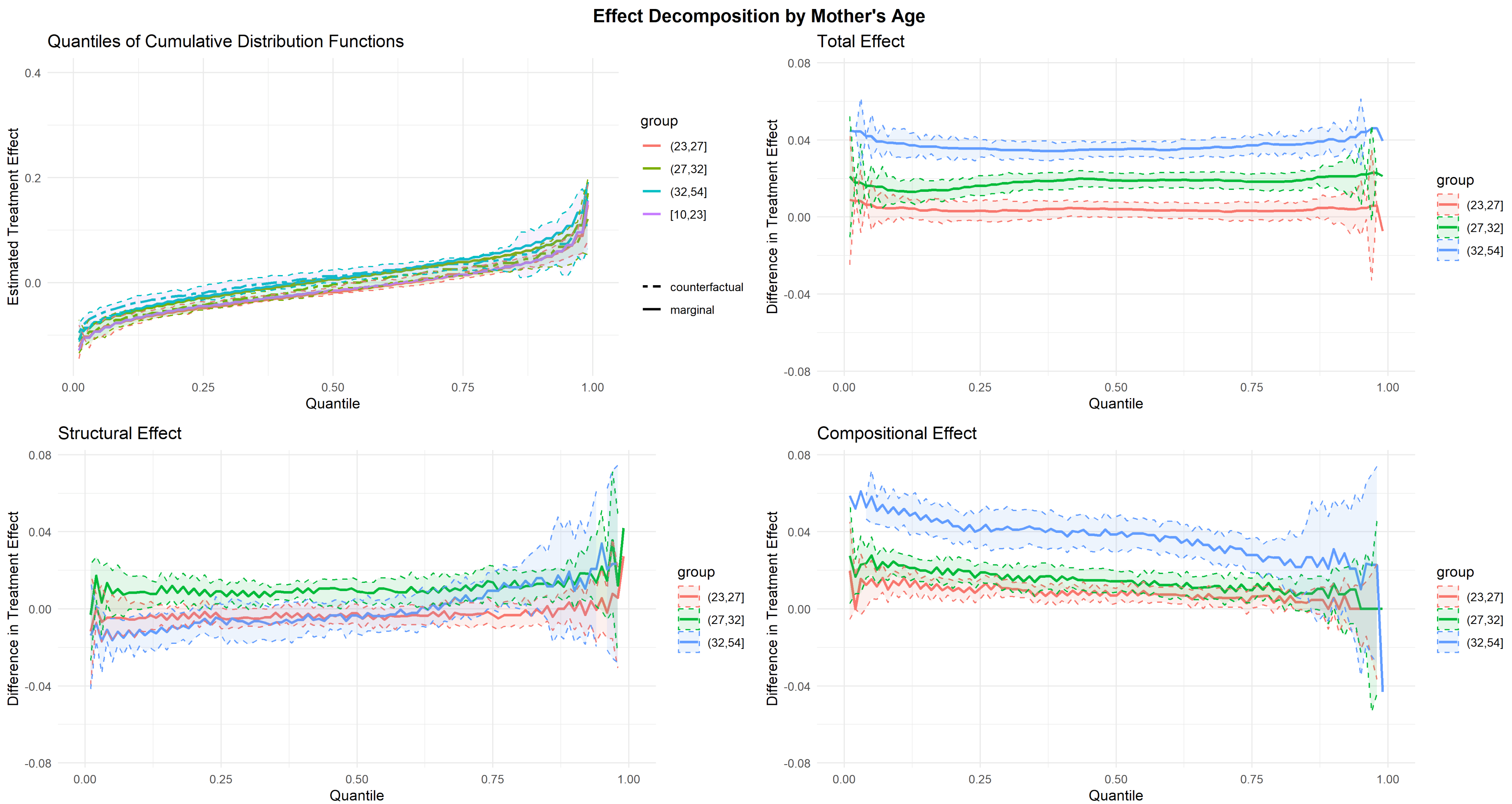

Figure 7: Standardized Birth Weight - Effect Decomposition by Mother's Age

Note: The figure displays the decomposition of the effect of a heat shock during pregnancy on standardized birth weight by mother's age. It shows the quantiles of the cumulative distribution function of each group and their counterfactual distribution (top left), the total effect difference (top right), structural effect (bottom left) and compositional effect (bottom right) difference between each group and the reference group. The counterfactual distribution and decomposition are derived with respect to the reference group which are mothers aged 10 to 23 (the lowest age quartile). Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in 11. Positive difference corresponds to effect mitigation with increasing age, whereas negative difference corresponds to effect amplification with increasing age respectively. Shaded areas show 95% confidence intervals.

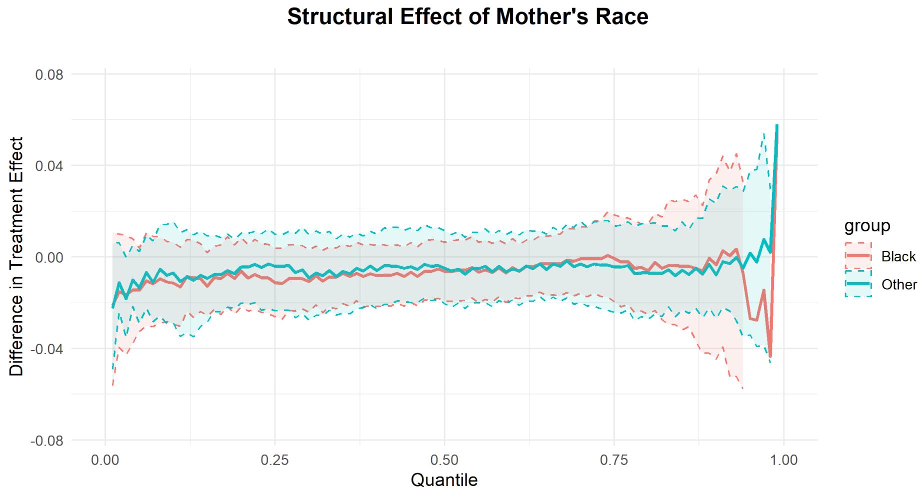

Figure 8: Standardized Birth Weight- Structural Effect of Mother's Race

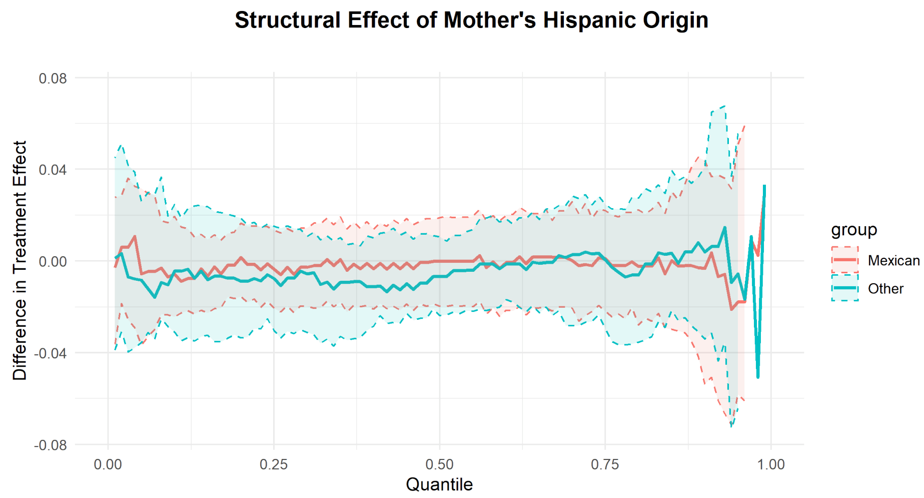

Figure 9: Standardized Birth Weight- Structural Effect of Mother's Hispanic Origin

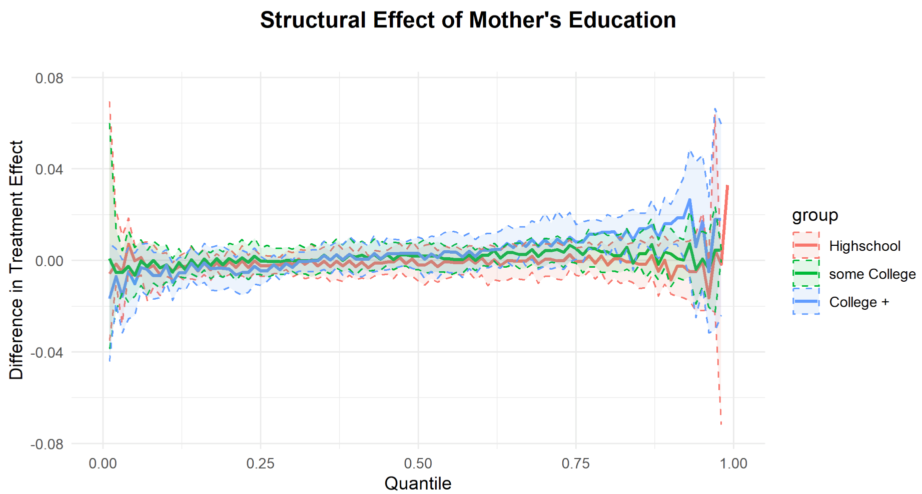

Figure 10: Standardized Birth Weight- Structural Effect of Mother's Education

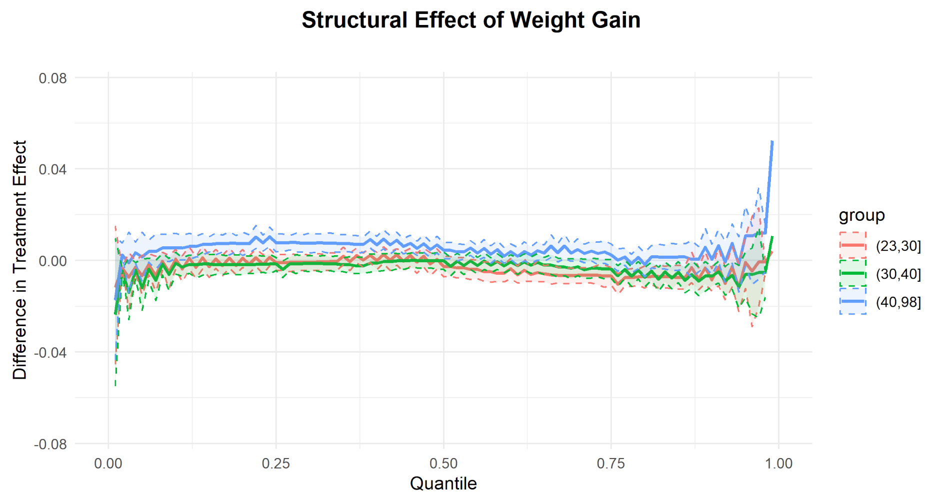

Figure 11: Standardized Birth Weight- Structural Effect of Weight Gain

Note: The figure shows the structural effect decomposition of the effect of a heat shock during pregnancy on standardized birth weight for different potentially modifying factors. Decomposition follows the procedure described in 1. Positive difference corresponds to effect mitigation, whereas negative difference corresponds to effect amplification respectively. Shaded areas show 95% confidence intervals.

4.3 Placebo Test

Table 7 presents the results of a "placebo test" in which we estimate the effect of a heat shock that occurred six months after birth. A heat shock after the date of birth is known to have no causal effect on birth outcomes. But if our estimates of the treatment effect of heat shocks during pregnancy only reflect trends in the birth outcomes or an omitted variable, we might as well see significant effects of the placebo shock. In addition, this serves as an indirect test to assess plausibility of the unconfoundedness assumption. Finding a significant effect of the heat shock happening after birth would make the unconfoundedness assumption less plausible.

Table 7: Placebo Test: Average Treatment Effect of shock 6 months after birth

	ATE	Std. Error
Standardized Birth Weight	-0.0049	0.0053
SGA	0.0012	0.0017

Notes: ATE is estimated using the inverse-propensity weighting in equation (8).
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

We find no significant average treatment effect of the heat shock 6 months after birth on either standardized birth weight or small for gestational age birth. As expected, the magnitude of effects is smaller than in the main analysis. Overall, the placebo test supports the effectiveness of our strategy in identifying causal impacts of heat shock exposure on health at birth.

5 Discussion

How does temperature exposure during pregnancy influence infants' health at birth? And who is most prone to climate change related health risks? This study evaluates the effects of temperatures and rain exposure on several outcomes for health at birth and study heterogeneity in these effects. It is the first to analyze heterogeneity in the effects of temperature shocks on infants' health at birth in a structured way and to shed a light on the possible mechanism of how these exposures work.

We show that there is strong heterogeneity in the effect of heat shocks on health at birth, which has not been studied in detail before. Analysis reveals that especially babies born to black, Mexican, and low-educated mothers are disproportionately prone to climate-related health risks. Given the combination of race, Hispanic origin, and lower education, it seems like disadvantaged populations are most prone to health risks of in-utero exposure to health at birth. A possible explanation might be the lack of protection. These mothers might not have the possibility to stay inside, since they might need to work outside and do not have access to air conditioning in their homes. Unfortunately, we do not have data available to explore this mechanism further. These racial health disparities have not been explored in detail in the literature, but given the diversity in the US population, it is imperative to understand how climate change impacts infants born to mothers of different racial and ethnic backgrounds. This can aid in identifying whom to target for the prevention of heat-associated morbidity.

For the fixed effects regression estimation approach, which is often used in the literature, we highlight problems arising from a mechanical correlation of gestation length and the exposure measures and discuss possible solutions. We find a significant effect on both sides of the temperature distribution, where cold temperatures positively affect standardized birth weight-related measures and gestation, whereas hot days negatively affect birth outcomes. Especially in the second and third trimesters, where the fetus grows most, the effects are strongest. Given the mechanical correlation between gestation length and the exposure, it is hard to gather insights into the effects of temperature exposure on gestation length. However, effects in the first and second trimesters, which are not prone to this mechanical correlation, show that extreme heat negatively affects gestation length.

There are however some limitations to our empirical approach. First, we only observe effects on live births in our data. Since the data does not include any information on miscarriage and stillbirth, we cannot measure the effects on most vulnerable infants, who did not survive up until the time of birth. Therefore, we expect our estimates for average effects to be a lower bound for the actual effect of temperature and weather exposure on health at birth. Second, we cannot control for mothers' mobility and actual temperature exposure. While we have information on the mother's county of residence, we cannot track the mother's mobility patterns. Additionally, we do not know whether she experienced any hot temperatures or if she was able to protect herself from extreme temperatures through heating, air conditioning, or location changes. From this perspective, our exposure measure is only a proxy for actual temperature and weather exposure. Third, the heterogeneity analysis unexpectedly reveals possible positive effects of heat wave exposure. As the literature agrees on negative effects of extreme heat on infants' health at birth and human health in general, we question these positive treatment effects found. This might point toward a possible adaptation the causal forest algorithm. In situations where the researcher has a strong prior about the effect to be estimated, solving the estimation problem under the constraint of negative treatment effects would be a possible extension of the algorithm.

Despite these limitations, this study provides important evidence of how temperature and extreme heat events affect health at birth and shed a light on heterogeneity in these effects. It introduces cutting-edge machine learning techniques to uncover heterogeneity in the effect of environmental factors on infant health and highlights the potential of machine learning approaches. Given the growing number of extreme weather events due to climate change and overall global warming, our results point out the most vulnerable groups which should be protected further from temperature and weather-related shocks.

References

Almond, Douglas and Janet Currie, "Killing Me Softly: The Fetal Origins Hypothesis," *Journal of Economic Perspectives*, September 2011, *25* (3), 153–72.

_ , Kenneth Y. Chay, and David S. Lee, "The Costs of Low Birth Weight," *The Quarterly Journal of Economics*, 08 2005, *120* (3), 1031–1083.

American College of Obstetricians and Gynecologists., "Committee Opinion No. 644: The Apgar Score.," *Obstetrics and gynecology*, Oct 2015, *126*, e52–e55.

Andalón, Mabel, João Pedro Azevedo, Carlos Rodríguez-Castelán, Viviane Sanfelice, and Daniel Valderrama-González, "Weather shocks and health at birth in Colombia," *World Development*, 2016, *82*, 69–82.

Athey, Susan, Julie Tibshirani, and Stefan Wager, "Generalized random forests," *The Annals of Statistics*, 2019, *47* (2), 1148 – 1178.

Bratti, Massimiliano, Prince Boakye Frimpong, and Simone Russo, "Prenatal Exposure to Heat Waves and Child Health in Sub-saharan Africa," IZA Discussion Papers 14424, Institute of Labor Economics (IZA) May 2021.

Breiman, Leo, "Random Forests," *Machine Learning*, Oct 2001, *45* (1), 5–32.

Chang, Grace, Marta Favara, and Rafael Novella, "The origins of cognitive skills and non-cognitive skills: The long-term effect of in-utero rainfall shocks in India," *Economics & Human Biology*, 2022, *44*, 101089.

Chen, Xi, Chih Ming Tan, Xiaobo Zhang, and Xin Zhang, "The effects of prenatal exposure to temperature extremes on birth outcomes: the case of China," *Journal of Population Economics*, 2020, *33*, 1263–1302.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 01 2018, *21* (1), C1–C68.

_ , Iván Fernández-Val, and Blaise Melly, "Inference on Counterfactual Distributions," *Econometrica*, 2013, *81* (6), 2205–2268.

_ , Mert Demirer, Esther Duflo, and Iván Fernández-Val, "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India," Working Paper 24678, National Bureau of Economic Research June 2018.

Currie, Janet and Maya Rossin-Slater, "Weathering the storm: Hurricanes and birth outcomes," *Journal of health economics*, 2013, *32* (3), 487–503.

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

_ , _ , and Jonathan Guryan, "Climate change and birth weight," *American Economic Review*, 2009, *99* (2), 211–217.

Hernández-Díaz, Sonia, Enrique F. Schisterman, and Miguel A. Hernán, "The Birth Weight "Paradox" Uncovered?," *American Journal of Epidemiology*, 08 2006, *164* (11), 1115–1120.

Imbens, Guido W., "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics," *Journal of Economic Literature*, December 2020, *58* (4), 1129–79.

Imbens, Guido W and Donald B Rubin, *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press, 2015.

Isen, Adam, Maya Rossin-Slater, and Reed Walker, "Relationship between season of birth, temperature exposure, and later life wellbeing," *Proceedings of the National Academy of Sciences*, 2017, *114* (51), 13447–13452.

Karlsson, Martin and Nicolas R Ziebarth, "Population health effects and health-related costs of extreme temperatures: Comprehensive evidence from Germany," *Journal of Environmental Economics and Management*, 2018, *91*, 93–117.

Kramer, Michael S., "Intrauterine Growth and Gestational Duration Determinants," *Pediatrics*, 1987, *80* (4), 502–511.

Le, Kien and My Nguyen, "The impacts of temperature shocks on birth weight in Vietnam," *Population and Development Review*, 2021, *47* (4), 1025–1047.

Molina, Oswaldo and Victor Saldarriaga, "The perils of climate change: In utero exposure to temperature variability and birth outcomes in the Andean region," *Economics & Human Biology*, 2017, *24*, 111–124.

National Center for Health Statistics, "Vital Statistics Natality Birth Data, 1989-2004. Public-use data file and documentation," Natality Micro Data 1989-2004.

North America Land Data Assimilation System, "Daily Air Temperatures and Heat Index, years 1979-2011," CDC WONDER Online Database, released 2012.

Robinson, P. M., "Root-N-Consistent Semiparametric Regression," *Econometrica*, 1988, *56* (4), 931–954.

Rocha, Rudi and Rodrigo R. Soares, "Water scarcity and birth outcomes in the Brazilian semiarid," *Journal of Development Economics*, 2015, *112*, 72–91.

Samuels, Louisa, Britt Nakstad, Nathalie Roos, Ana Bonell, Matthew Chersich, George Havenith, Stanley Luchters, Louise-Tina Day, Jane E. Hirst, Tanya Singh, Kirsty Elliott-Sale, Robyn Hetem, Cherie Part, Shobna Sawry, Jean Le Roux, and Sari Kovats, "Physiological mechanisms of the impact of heat during pregnancy and the clinical implications: review of the evidence from an expert group meeting," *International Journal of Biometeorology*, 2022, *66* (8), 1505–1513.

Stürmer, Til, Kenneth J. Rothman, Jerry Avorn, and Robert J. Glynn, "Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution–a simulation study.," *American journal of epidemiology*, Oct 2010, *172*, 843–54.

Wilde, Joshua, Bénédicte H Apouey, and Toni Jung, "The effect of ambient temperature shocks during conception and early pregnancy on later life outcomes," *European Economic Review*, 2017, *97*, 87–107.

World Health Organization, *COP24 special report: health and climate change*, World Health Organization, 2018.

Zenzes, Johanna M., "Uncovering Sources of Heterogeneity in the Effects of Maternal Smoking on Infants' Health at Birth," `https://drive.google.com/file/d/1qPI7wVMaSoXTV8Udx56kWlAEB9WESdwn/view` 2022. Accessed: 2023-02-13.

Zhang, Xin, Yixuan Wang, Xi Chen, and Xun Zhang, "Associations between prenatal sunshine exposure and birth outcomes in China," *Science of The Total Environment*, 2020, *713*, 136472.

Appendix

A Weather Data Overview

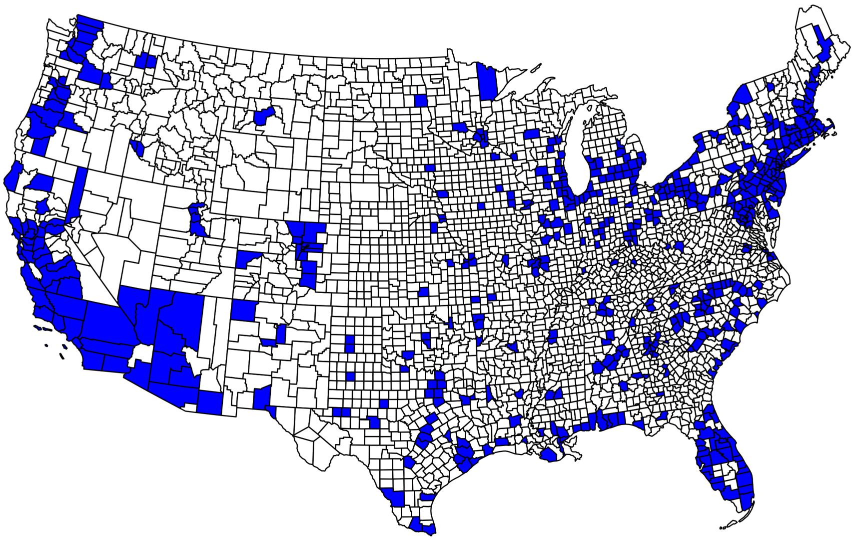

Figure A.1: Counties with a population larger than 100,000 included in the analysis

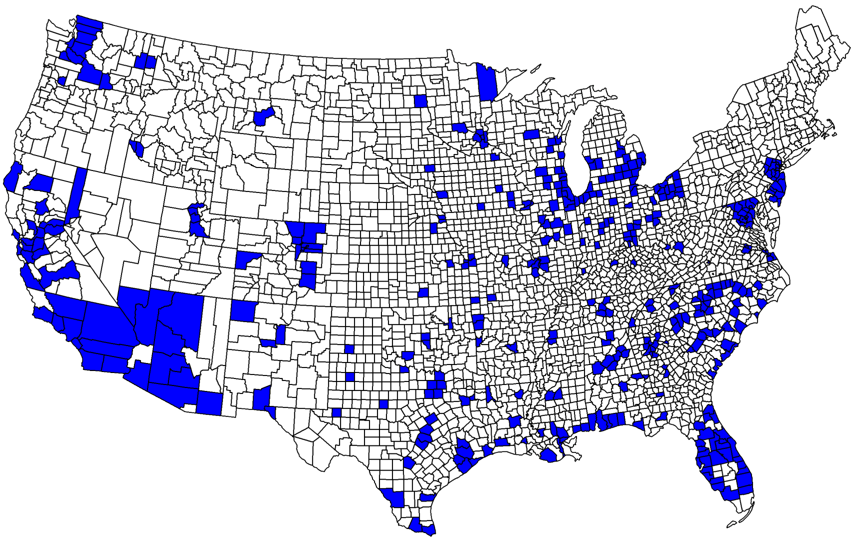

Figure A.2: Counties with sufficient overlap in propensity score estimates

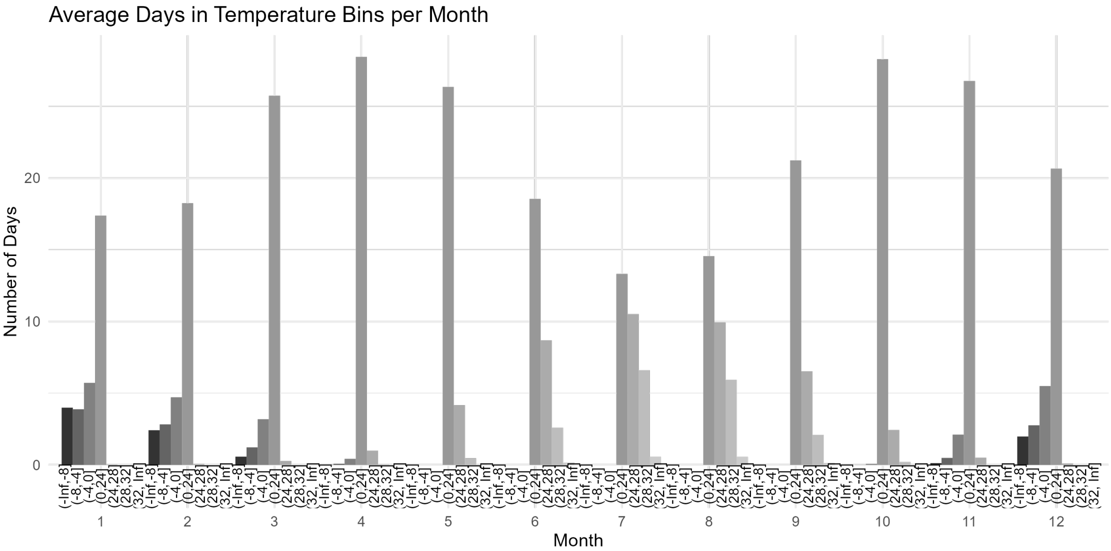

Figure A.3: Average count of days in temperature bins per month

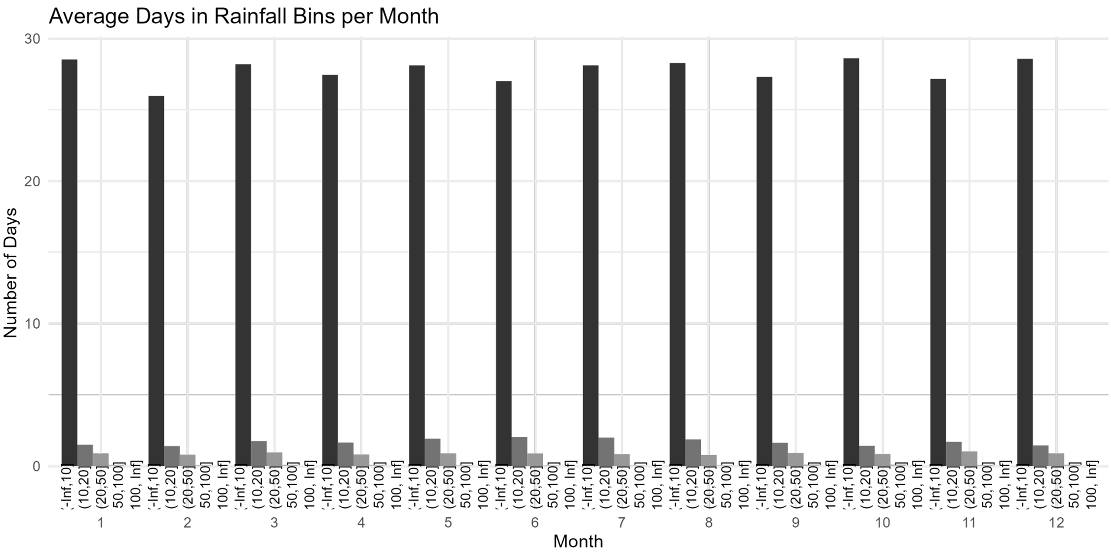

Figure A.4: Average count of days in rainfall bins per month

B Mechanical Correlation of Gestation and Temperature Exposure

Table A.1 turns to an illustration of the mechanical correlation between gestation length and temperature exposure and the resulting problems with estimated effects. The estimates show effect patterns for the first and second trimesters, which are in line with what we found for the standardized outcomes. However, turning to the effects in the third trimester, we observe very large positive effects. The large positive effects arise from the mechanical correlation between gestation length and exposure to temperature. These mainly measure the effect of an additional day of the pregnancy rather than the actual effect of temperature exposure. This is because not all pregnancies observed last 9 months. We might therefore compare very lightweight preterm babies to full-term babies, whose count in each temperature bin is most likely to be larger, given a longer gestation length. Therefore, we see large positive effects of the exposure, rather than the expected smaller effects.

C Decomposition

Algorithm 1: Effect Decomposition

Repeat multiple times using different random splits into auxiliary and main sample:

- Split the sample into auxiliary (A) and main sample (M)
- On Auxiliary sample: train the causal forest
- On Main sample:
  - Obtain estimates $\hat{\tau}(x)$ of the treatment effect via the causal forest trained on A
  - Define number of populations $k \in \mathcal{K}$ based on the partitioning variable $X^* \in X$ (continuous variables need to be categorized)
  - $\forall k \in \mathcal{K} \setminus \{0\}$: decompose total difference between the treatment effect in each group k and the reference group 0 into structural and compositional effect
    - Obtain estimates $\hat{F}_{X_k}$ of the covariate distributions F_{X_k} via $\hat{F}_{X_k}(x) = \frac{1}{n_k}\sum_{i=1}^{n_k} 1_{\{X_{ki} \leq x\}}, \quad k \in \mathcal{K}$
    - Obtain estimates $\hat{F}_{\tau(x)^{(0)}|X_0}$ of the conditional distribution via $F_{Y|X}(y|x) = \Lambda(P(x)'\beta(y)), \quad \forall y \in \mathcal{Y}$
    - Obtain estimates of the counterfactual distributions via $F_{\tau\langle j|k\rangle}(y) := \int_{\mathcal{X}_k} F_{\tau_j|X_j}(y|x)dF_{X_k}(x)$
    - Obtain decomposition via

$$\underbrace{F_{\tau\langle k|k\rangle}(y) - F_{\tau\langle j|j\rangle}(y)}_{\textit{Total Effect}} = \underbrace{[F_{\tau\langle k|k\rangle}(y) - F_{\tau\langle j|k\rangle}(y)]}_{\textit{Structural Effect}} + \underbrace{[F_{\tau\langle j|k\rangle}(y) - F_{\tau\langle j|j\rangle}(y)]}_{\textit{Compositional Effect}}$$

Note: This effect decomposition algorithm can also be used for quantiles by using following transformation $Q_{\tau\langle j|k\rangle}(y) := F^{-1}_{\tau\langle j|k\rangle}(p), \;\; p \in (0,1)$.

D Temperature Effects - Trimester

We further analyze sensitivity to the timing of the temperature exposure. Table A.2 shows the effects separately for each trimester, following equation (13). Results are not sensitive to including additional controls such as weather and mother's fixed effects, see **??**. For standardized birth weight (columns (1), (2), (3)), we observe an increase in effect size with increasing pregnancy length. The effects of exposure to hot temperatures are most harmful in the second and third trimester. For the effects of cold exposure, we do only observe an increase in the third trimester. Regarding SGA, there is no strong effect in either the first or the second trimester. Significant effects are concentrated in the third trimester. Effects of temperature exposure on standardized Apgar Score are mostly not significant, exceptions are an increase in Apgar score found for extreme heat in the first and second trimester and a decrease for mild warm temperatures between 24-28°C in the third trimester.

Table A.1: Trimester Temperature Effects: Mechanical Correlation

Dependent Variables:	Birth Weight			Gestation Length		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
TR1: Avg. Temp. < -8°C	0.0170			-0.0001		
	(0.0450)			(0.0003)		
TR1: Avg. Temp. -8 - -4 °C	0.0774			0.0000		
	(0.0513)			(0.0003)		
TR1: Avg. Temp. -4 - 0 °C	0.0308			-0.0005**		
	(0.0354)			(0.0002)		
TR1: Avg. Temp. 24 - 28 °C	-0.0045			-0.0004**		
	(0.0310)			(0.0002)		
TR1: Avg. Temp. 28 - 32 °C	-0.0075			-0.0008*		
	(0.0451)			(0.0004)		
TR1: Avg. Temp. ¿32 °C	-0.1618			-0.0001		
	(0.1158)			(0.0008)		
TR2: Avg. Temp. < -8°C		-0.0164			0.0004	
		(0.0520)			(0.0002)	
TR2: Avg. Temp. -8 - -4 °C		0.0151			-0.0004	
		(0.0561)			(0.0003)	
TR2: Avg. Temp. -4 - 0 °C		-0.1307***			-0.0004*	
		(0.0418)			(0.0002)	
TR2: Avg. Temp. 24 - 28 °C		-0.1174***			-0.0006***	
		(0.0259)			(0.0002)	
TR2: Avg. Temp. 28 - 32 °C		-0.1304***			-0.0003	
		(0.0415)			(0.0003)	
TR2: Avg. Temp. $>$ 32 °C		-0.4329***			-0.0010	
		(0.0633)			(0.0008)	
TR3: Avg. Temp. < -8°C			7.9138***			0.0770***
			(0.3602)			(0.0037)
TR3: Avg. Temp. -8 - -4 °C			5.2416***			0.0483***
			(0.2737)			(0.0026)
TR3: Avg. Temp. -4 - 0 °C			10.9569***			0.1043***
			(0.3304)			(0.0035)
TR3: Avg. Temp. 24 - 28 °C			7.9763***			0.0817***
			(0.1633)			(0.0017)
TR3: Avg. Temp. 28 - 32 °C			6.1534***			0.0646***
			(0.2224)			(0.0025)
TR3: Avg. Temp. $>$ 32 °C			6.7083***			0.0743***
			(0.2554)			(0.0028)
Weather Controls	Yes	Yes	Yes	Yes	Yes	Yes
Mother's Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Fixed-effects						
Year of Birth	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth	Yes	Yes	Yes	Yes	Yes	Yes
County	Yes	Yes	Yes	Yes	Yes	Yes
State	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth-County	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704
R^2	0.1105	0.1105	0.1295	0.0402	0.0402	0.1841
Within R^2	0.1002	0.1002	0.1195	0.0295	0.0295	0.1750

Notes: Entries show the coefficient on the relevant temperature exposure measure. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension. Standard errors clustered by County of residence in parentheses.

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table A.2: Trimester Temperature Effects

Dependent Variables:	Standardized Birth Weight			SGA			Standardized Apgar Score		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
TR1: Avg. Temp. ¡-8°C	0.0694			0.0404			-0.3315		
	(0.0931)			(0.0275)			(0.2334)		
TR1: Avg. Temp. -8 - -4 °C	0.1546			-0.0261			0.3827		
	(0.1038)			(0.0266)			(0.2754)		
TR1: Avg. Temp. -4 - 0 °C	0.1807**			-0.0078			-0.0037		
	(0.0725)			(0.0189)			(0.1486)		
TR1: Avg. Temp. 24 - 28 °C	0.0887*			-0.0204			0.0388		
	(0.0538)			(0.0140)			(0.0708)		
TR1: Avg. Temp. 28 - 32 °C	0.0902			-0.0346**			-0.0054		
	(0.0605)			(0.0171)			(0.0939)		
TR1: Avg. Temp. $>$ 32 °C	-0.2429*			0.0637***			0.4349***		
	(0.1288)			(0.0231)			(0.1571)		
TR2: Avg. Temp. < -8°C		-0.0979			0.0330			-0.0808	
		(0.1148)			(0.0302)			(0.2343)	
TR2: Avg. Temp. -8 - -4 °C		0.0858			-0.0834**			0.2481	
		(0.1219)			(0.0335)			(0.2775)	
TR2: Avg. Temp. -4 - 0 °C		-0.1895**			0.0302			0.0661	
		(0.0860)			(0.0222)			(0.1641)	
TR2: Avg. Temp. 24 - 28 °C		-0.0815			-0.0021			-0.0992	
		(0.0530)			(0.0139)			(0.0733)	
TR2: Avg. Temp. 28 - 32 °C		-0.1102*			0.0183			-0.0633	
		(0.0609)			(0.0200)			(0.0967)	
TR2: Avg. Temp. $>$ 32 °C		-0.5523***			0.0612***			0.5212***	
		(0.0880)			(0.0176)			(0.1589)	
TR3: Avg. Temp. < -8°C			0.3122			-0.0691			-0.2026
			(0.2270)			(0.0481)			(0.2616)
TR3: Avg. Temp. -8 - -4 °C			0.6775***			-0.1600***			0.3490
			(0.1902)			(0.0435)			(0.2903)
TR3: Avg. Temp. -4 - 0 °C			0.8087***			-0.1697***			0.2743
			(0.2186)			(0.0413)			(0.1766)
TR3: Avg. Temp. 24 - 28 °C			-0.8682***			0.1283***			-0.2701***
			(0.1379)			(0.0258)			(0.0947)
TR3: Avg. Temp. 28 - 32 °C			-1.0127***			0.1313***			0.0743
			(0.1579)			(0.0267)			(0.1979)
TR3: Avg. Temp. $>$ 32 °C			-1.8491***			0.2694***			-0.1215
			(0.0852)			(0.0533)			(0.2480)
Weather Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mother's Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed-effects									
Year of Birth	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth-County	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics									
Observations	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704
R^2	0.0900	0.0900	0.0901	0.0376	0.0376	0.0376	0.0283	0.0283	0.0283
Within R^2	0.0815	0.0815	0.0816	0.0338	0.0338	0.0338	0.0025	0.0025	0.0025

Notes: Entries show the coefficient on the relevant temperature exposure measure, scaled by 1000 for ease of reading. This means that a one-unit change in *Avg. Temp. <-8°C* is associated with a 0.0004 unit change in standardized birth weight in column (1), holding all other variables constant. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension.
Standard errors clustered by County of residence in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

E Rainfall Effects

We additionally analyze effects of rainfall exposure. Table A.3 shows the results of equation (12) for exposure to rainfall using our preferred specification including mother's characteristics and weather controls. Milder rainfall of 10 to 50 mm per day shows slightly positive effects compared to exposure to rainfall between 0 and 10mm. Exposure to very extreme rainfall shows a weakly significant reduction of 0.0034 which corresponds to a reduction of 1.63 grams on average. Including additional controls (column (2)), none of the estimated effects remains significant. Similarly, we cannot observe any significant effect of rainfall on SGA birth (column (4)), indicating that most vulnerable infants are not affected by rainfall. For the standardized Apgar score, we see a significant reduction in exposure to extreme rainfall. Similarly, we cannot find significant effects of any rainfall bin when analyzing trimester-specific settings as specified in equation (13). See table A.4 for effect estimates for rainfall exposure. We therefore cannot find significant effects of rainfall exposure on any of the outcomes of health at birth. No effects of rainfall on health at birth is partly in line with previous literature. Andalón et al. (2016) find no significant effect of either drought or heavy rainfall on birth weight ad gestation length. In contrast, Rocha and Soares (2015) find positive effects of positive rainfall shocks during pregnancy on birth weight. They however study birth in rural Brazil, where no rainfall easily leads to water scarcity.

Table A.3: Rainfall Effects

Dependent Variables:	Standardized Birth Weight		SGA		Standardized Apgar Score	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Rain 10-20mm	0.6208***	0.1173	-0.1530***	-0.0297	0.0259	0.0183
	(0.1122)	(0.1382)	(0.0230)	(0.0292)	(0.1942)	(0.1956)
Rain 20-50mm	0.3631***	-0.1776	-0.1140***	0.0033	-1.0747***	-1.0087***
	(0.1347)	(0.1677)	(0.0311)	(0.0405)	(0.3631)	(0.3666)
Rain 50-1000mm	0.3772	0.1585	-0.1329	-0.1018	2.0544	2.4154*
	(0.4733)	(0.4463)	(0.1186)	(0.1057)	(1.3634)	(1.4536)
Rain +100mm	-3.3764*	-2.0670	0.4401	0.0979	-12.3021***	-12.0226**
	(1.9932)	(2.2034)	(0.5869)	(0.6275)	(4.2158)	(4.8551)
Weather Controls		Yes		Yes		Yes
Mother's Characteristics		Yes		Yes		Yes
Fixed-effects						
Year of Birth	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth	Yes	Yes	Yes	Yes	Yes	Yes
County	Yes	Yes	Yes	Yes	Yes	Yes
State	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth-County	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	19,865,677	18,595,704	19,865,677	18,595,704	19,865,677	18,595,704
R^2	0.0091	0.0900	0.0039	0.0376	0.0265	0.0283
Within R^2	0.0000	0.0815	0.0000	0.0338	0.0000	0.0025

Notes: Entries show the coefficient on the relevant rainfall exposure measure, scaled by 1000 for ease of reading. This means that a one-unit change in *Rain 10-20mm* is associated with a 0.0006 unit change in standardized birth weight in column (1), holding all other variables constant. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension.

Standard errors clustered by County of residence in parentheses.

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table A.4: Trimester Rainfall Effects

Dependent Variables:	Standardized Birth Weight			SGA			Standardized Apgar Score		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
1st Trimester: Rain 10-20mm	0.1488			-0.0309			0.0202		
	(0.1178)			(0.0316)			(0.2056)		
1st Trimester: Rain 20-50mm	-0.0833			0.0496			-0.9311^{***}		
	(0.1713)			(0.0459)			(0.3505)		
1st Trimester: Rain 50-1000mm	-0.1022			0.0017			2.1038		
	(0.5296)			(0.1514)			(1.5005)		
1st Trimester: Rain +100mm	0.6805			-0.0834			-16.5432^{***}		
	(2.8214)			(0.7782)			(5.2608)		
2nd Trimester: Rain 10-20mm		0.0077			0.0217			-0.0046	
		(0.1212)			(0.0359)			(0.2304)	
2nd Trimester: Rain 20-50mm		-0.2106			-0.0070			-0.6940^{*}	
		(0.1840)			(0.0469)			(0.3841)	
2nd Trimester: Rain 50-1000mm		0.2515			-0.0812			2.5004^{*}	
		(0.6458)			(0.1660)			(1.4930)	
2nd Trimester: Rain +100mm		-2.8722			-0.4605			-9.2522^{*}	
		(2.4008)			(0.7331)			(4.9711)	
3rd Trimester: Rain 10-20mm			0.1990			-0.0765			0.1055
			(0.3400)			(0.0662)			(0.2659)
3rd Trimester: Rain 20-50mm			-0.3037			-0.0542			-1.4165^{***}
			(0.3140)			(0.0799)			(0.4768)
3rd Trimester: Rain 50-1000mm			0.4325			-0.3017			2.6743
			(0.8256)			(0.1933)			(1.6302)
3rd Trimester: Rain +100mm			-5.3092			1.1469			-5.6922
			(3.4064)			(1.1729)			(5.2821)
Weather Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mother's Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed-effects									
Year of Birth	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month of Birth-County	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics									
Observations	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704	18,595,704
R^2	0.0900	0.0900	0.0900	0.0376	0.0376	0.0376	0.0283	0.0283	0.0283
Within R^2	0.0815	0.0815	0.0815	0.0338	0.0338	0.0338	0.0025	0.0025	0.0025

Notes: Entries show the coefficient on the relevant rainfall exposure measure, scaled by 1000 for ease of reading. This means that a one-unit change in *Rain 10-20mm* is associated with a 0.0001 unit change in standardized birth weight in column (1), holding all other variables constant. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension.
Standard errors clustered by County of residence in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*