THE UNIVERSITY OF CHICAGO

ESSAYS ON ENVIRONMENTAL POLICIES IN CHINA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE IRVING B. HARRIS
GRADUATE SCHOOL OF PUBLIC POLICY STUDIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
YUQI SONG

CHICAGO, ILLINOIS
JUNE 2023

Contents

List of Tables

List of Figures

ACKNOWLEDGMENTS

I would like to express sincere thanks to my doctoral advisors and dissertation committee members, Professor Ryan Kellogg, Professor Michael Greenstone, and Professor Amir Jina. I want to thank Professor Kellogg for his unceasing support and guidance since the very first day of my PhD at Harris, Professor Greenstone for his rigorous advising and encouragement for all my research projects since undergraduate, and Professor Jina for providing every detailed comment and helpful research tip when I develop challenging and uncertain proposals. I have benefited greatly from my committee, learning to be a constructive academic researcher in environmental and energy economics, and I want to dedicate the greatest gratitude to them for being my guide throughout my academic career.

Secondly, I am grateful to all faculties, fellow students, co-authors, and administrative staffs from the Harris School of Public Policy, the Energy Policy Institute at the University of Chicago (EPIC), Kenneth C. Griffin Department of Economics, Booth School of Business, the National Research Traineeship (NRT) Program in Computational Data Science to Advance Research at the Energy-Environment Nexus at the University of Chicago, the Research Computing Center at the University of Chicago, and MIT Sustainability and Urbanization Lab, for all their supports in data, coding, presentation, and general research and career

advises throughout the course of my PhD. I have had the most wonderful and memorable experience studying in the community composed of these people at the University of Chicago.

I also want to thank the many Chinese economists who have read and commented on my research, including my co-authors. It is greatly meaningful that my works are encouraged by people familiar with and concerned about the local environmental policies of the country.

Lastly, I would like to express special gratitude to my family and friends who have helped me in both research and life. I could not have continued this far without their love and support.

ABSTRACT

As the world's largest developing economy with growing concerns about how to counter environmental and climate shocks while maintaining its economic growth, China has developed various environmental policy tools throughout the past few decades. My dissertation includes three chapters of research that study the socio-economic impacts of two important Chinese environmental policies, the public good of quality weather forecasts, which aims to help people's adaptation to extreme weathers happening in the near future, and the long-run road rationing policy applied in nine major Chinese cities, which aims to limit emissions from vehicles on road and lower city level pollutions. Overall, my research identifies the differential policy impacts of these different policy tools of China in tackling environmental problems.

The first two chapters are research under the greater project of "The Value of Weather Forecasts". For this project, I construct a novel dataset of 24-hour city-level weather forecasts in China, using Google speech-to-text API to transcribe videos of the national weather forecast TV programs to collect the actual information broadcast and received by the general public. From these research, I find that there exist significant behavior responses to the accuracy of weather forecast information in China. In Chapter 1, I show that accurate instead of inaccurate daily temperature forecasts of uncomfortable temperatures (extreme

hot and medium-cold) lead to significant decreases in labor working hours per day. This shows that accurate weather forecast information helps in laborer's decisions to work less under weather shocks, in order to avoid potential health risks. Correspondingly, improved accuracy of weather forecasts contributes significant social values. The welfare analysis of this chapter estimates a marginal value of weather forecast accuracy as 930 2015 Yuan (148 2015 USD) per worker per year. Social benefits of accurate weather forecasts are also represented in Chapter 2, which demonstrates that when realized temperatures are extremely low, the negatively impacted average social sentiment (summarized with natural language processing analysis of city-level daily social media posts in 2014) is significantly improved, if accurate instantaneous temperature forecasts are provided.

The third chapter analyzes the impacts of a well-known environmental policy in China, the end-number license plate policy, of a long-run, less strict version imposed over a set of 9 big cities of China restricting one-fifth of private vehicles per day on weekdays. This version of the road rationing policy is shown to have limited impacts in effectively reducing city-level air pollutions over the timeline of a decade, contrary to previous literatures showing that the short-run, strict version of this policy can significantly improve air qualities. The ambiguity of the policy effects implies people's behavior changes in response to the long-run road rationing policy, and provides useful implications on motivating instead of requiring people to change their daily activities for the society goal of cutting air pollutant emissions with different policy tools.

Chapter 1

The Value of Weather Forecasts: Labor Responses to Accurate and Inaccurate Temperature Forecasts in China

Abstract

This paper evaluates the economic value of the accurate information of weather forecasts, a common and popular public good continuously invested by the government in modern societies. Labor decisions of hours worked per day are found to respond to day-ahead temperature forecasts, and only forecasts that are perceived as accurate are incorporated. From this, improvements in weather forecast accuracy generate large social benefits. With the setting in China, where the developing economy provides a trusted uniform source of national weather forecasts to its large population, I collect a novel dataset of the city-level day-ahead weather forecasts directly broadcast to the general public through video transcription for over 2000 days of the country's popular weather forecasts TV program. Constructing the metric of perceived forecast accuracy based on the medium-run root mean squared errors of historical temperature forecasts, I run a regression with the interaction of forecasts and forecast accuracy to estimate how labor response to forecast temperatures varies under different levels of forecast accuracy across time and space. My main regression results suggest large reductions of labor supply up to 4.5 and 1.2 hours per day under hot (above $30C$) and medium-cold ($15C - 25C$) daily maximum temperature forecasts only when these forecasts are accurate. Instead, when forecasts are inaccurate, those negative labor responses diminish. Using a simple utility maximization model for the next-day labor decision and evaluating with the regression estimates, I estimate a large marginal value of forecast accuracy of 930 2015 Yuan (148 USD) per capita per year. For the entire country, an average 3.9% increase in forecast accuracy from 2011 to 2015 generates a partial social benefit from the labor sector alone at 25.3 billion 2015 Yuan (4.03 billion USD) per year, about covering the annual cost of the national weather forecasting system.

1.1 Introduction

Human beings have been interested in weather forecasting since thousands of years ago, when meteorology, the study of weather, has been founded. Over the past hundred years, developments in technology and modern meteorology have led to significant increase of weather forecast accuracy (Shuman, 1989)[1]. Into the 21st century, weather forecasting is expected to continue progressing at an even greater pace (Teague and Gallicchio, 2017).

The economic value of weather forecasts is of great interest to many from physical scientists to policy makers (Murphy, 1993). For one thing, accurate weather forecast information is important to people's everyday decision makings. People are using these information to efficiently plan their daily activities ahead of time such that potential damages caused by future weather shocks can be avoided (Katz and Murphy, 2005; Guido etal., 2021). For another, developing accurate weather forecasts is costly. Every year, modern societies are spending billions on national and international forecasting systems to ensure that the most reliable weather forecasts can be delivered to the general public. Therefore, how weather forecasts have affected people's decisions and what social benefit is generated with more accurate weather forecasts are important questions to answer.

In this paper, I identify the impacts of accurate weather forecasts on the decision-making

[1]There are multiple metrics meteorologists use to describe weather forecast accuracy, including absolute mean errors, forecast skills and threat scores. In percentage terms, the recorded improvement of weather forecast accuracy is large over the past decades across different metrics. Shuman (1989) shows the error score for 36-hour predictions of geopotential height at 500mb drop by more than 50% from 1955 to 1988. NOAA shows the NCEP forecast skill for 500mb geopotential height increases to more than 3.6 times from 1955 to 2015, and the mean absolute errors of short-range (3-7 days) maximum temperature forecasts drops by about 45% from 1972 to 2017 (source: https://www.wpc.ncep.noaa.gov/).

of the Chinese labor population, and estimate a large partial value of accurate weather forecasts in China. Showing that labors respond with decreasing hours worked per day when they are provided with accurate forecasts for uncomfortable temperatures, I demonstrate that the information of weather forecasts is being incorporated in people's decisions on labor. Finding that people reduce their labor supply under accurate instead of inaccurate weather forecasts, I deduce that people make use of historical forecast accuracy in order to avoid judgment errors in labor decision-making under mis-information about future weathers. Through welfare analysis, I address that accurate weather forecasts are important to the society because they contribute large social benefits enough to cover the expenditures developing the forecasts. Furthermore, my research provides evidence for the value of information in shaping people's adaptation behavior to climate shocks, and my approach will be useful in valuing similar public goods closely related to climate and technology development.

I select the research setting in China where the largest developing economy has a great labor population exposed under climate risks. In the meantime, heterogeneous workers have been used to the same sourced free-of-charge national weather forecasts because of the country's vigorous investments into the provision of this information as a public good. To access the actual weather forecasts perceived by the Chinese population every day, I collect a novel dataset of city level day-ahead weather forecasts that has not been storage directly. I transcribe over 130 hours of video recordings of the nation's popular weather forecasts program aired every evening on TV with Google speech-to-text API, and gather a panel dataset of the 24-hour ahead temperature and precipitation category forecasts for cities all across the country over 2000 days since 2010.

For empirical analysis, I explore the impacts of these day-ahead weather forecasts on labor supply, represented with the individual reported number of hours working from China Health and Nutrition Survey (CHNS) by week and city, 2011 and 2015. Labor decisions on hours worked per period are important adaptation behavior to climate shocks, because labors are motivated to work less under uncomfortable temperatures in order to avoid disutilities likely from health risks (Graff Zivin and Niedell, 2014; Garg, Gibson and Sun, 2020; Rode etal., 2022). To separate the impacts of forecasts from those of realized temperatures in the simple regression of labor hours on a non-linear function of daily maximum temperature forecasts similar to previous literature, I consider the labor adaptation to forecast information relevant to its perceived accuracy. Adding to the regression an interaction term of forecast function and historical forecast accuracy, I apply an empirical design similar Carleton etal. (2020) to estimate the differential labor responses under space-time varying historical accuracy of weather forecasts, measured by the medium-run root-mean-squared error ($RMSE$) of daily maximum temperature forecasts over the previous half-year rolling window.

My main results of the baseline regression with forecasts and forecast accuracy interaction show that accurate forecasts of extreme hot (above $30C$) and medium-cold ($15C - 25C$) temperatures lead to large magnitudes of labor reduction (up to 4.5 hours for hot and up to 1.2 hours for medium-cold). In the meantime, inaccurate forecasts induce no such responses in labor. Therefore, I find that accurate weather forecasts can introduce early avoidance behaviors in labor such that people work less under uncomfortable temperature forecasts. These baseline results are verified by various robustness checks, and heterogeneity analysis

shows that results are more prominent among cities with less heating availability and individuals more vulnerable in health and economic backgrounds.

Lastly, I construct a single-period theoretical framework outlining workers' choice of next-day labor supply based on temperature forecasts and the perceived accuracy of such information. With this model and the main regression estimates, I quantitatively estimate the partial value of accurate weather forecasts from the behavioral response in labor sector in China. With implications of exaggerated beliefs under inaccurate forecasts, my method conveniently identifies a quadratic utility function for the decision makers with an impact function of realized weathers estimated by the labor response under sample maximum forecast accuracy, and a scale factor determined by the labor supply elasticity referenced from Chinese labor literature. In the end, I calculate that the marginal value of weather forecast accuracy (represented by $RMSE$ of daily maximum temperature forecasts) brings about 930 2015 Yuan (148 USD) for an average worker in China every year. In total, an average 3.9% increase of city level annual forecast accuracy from 2011 to 2015 generates a large social benefit of 25.3 billion 2015 Yuan (4.03 billion USD) per year for almost the entire country from labor sector alone, about covering the annual spending of the national weather forecasting system.

This paper contributes to literatures studying the impacts of weather and climate forecasts on economic activities. While existing literature focuses on the medium-range month-ahead probabilistic forecasts useful for longer-run production decisions in specific industries, this paper extend the impacts of study to short-range next-day weather forecasts more com-

monly known to the general labor force and instead explores their short-run decision making of labor take-up with forecast information. Among these literatures, Rosenzweig and Udry (2019) finds that only skillful weather forecasts could affect farmers' decision makings in rural India. Shrader (2020) discovers that the production in fishery responds to medium run ENSO forecasts. Downey, Lind and Shrader (2021) estimates a value of monthly rainfall level predicted by ENSO events because the information provides employment benefits to firms.

This paper also contributes to the broader literatures accessing the value of information. Specifically, this paper estimates the social benefit for a type of scientific forecasting which usually requires greater amount and longer run financial supports. More literatures are for financial and economic forecasts (Ivkovic and Jegadeesh, 2004; Goodarzi, Perera and Bunn, 2019) or the value of real-time information in environmental economics (Barwick, Li, Lin and Zou, 2019). Among the smaller volume of literatures on meteorological forecasts specifically, both the existing forecasting system and any marginal improvement of forecast accuracy have been estimated with large values (Katz and Murphy, 2005). Survey-based contingent valuation method is one common approach in these literatures. The Chinese survey in 2006 estimates the country's weather forecasting system to be a large social value of 46.5 billion Yuan in 2006 (5.83 billion USD) (Yuan, Sun and Wang, 2016). Similarly, the US national survey in 2006 estimates a 31.5 billion USD benefit of weather forecasts (Lazo, Morss and Demuth, 2009).

This paper estimates a large monetary value of accurate weather forecasts as other lit-

eratures. Similarly analyzing the short-range weather forecasts with revealed preference, Bakkensen, Lemoine and Shrader (2022) finds the increase of the mortality rate under greater forecast errors in the US that implies a large value of 75.1 billion USD per unit decrease in forecast error standard deviation. Other related research also quantifies the large value of weather forecasts through modeling and case studies, but concentrated more on developed economies. Nurmi etal. (2012) considers the forecast impacts on the transportation sector in the European Union and estimates a lower bound value of 3.4 billion Euros (4.5 billion 2012 USD) for 100% accurate forecasts. Fox, Turner and Gillespie (1999) studies the impacts of precipitation forecasts on agriculture production in Canada and translate the value of these forecasts to an average 100 CDN (67.6 1999 USD) per hectare per year. The less common but critical disaster forecasts are also studied to have large social values. Martinez (2020) shows that increased accuracy of hurricane forecasts in the US since 1970 leads to a total reduction in damage worth 82 billion USD, far exceeding the amount invested by the government.

This paper will proceed as the following. Section 2 describes the data and summary statistics. Section 3 outlines the empirical design. Section 4 presents the main results, heterogeneity analysis and robustness checks. Section 5 discusses the valuation model and results. Section 6 concludes.

1.2 Data

1.2.1 Weather Forecast Data

In this paper, I choose to study the 24-hour ahead daily and city level weather forecasts in China. This forecast product is selected for several reasons. Firstly, the nationwide forecasts are produced by China Meteorological Administration (CMA)[2], usually regarded as the authority and is the almost uniform source of forecast information in the country for decades. Secondly, the day-ahead weather forecasts are normally one of the most important forecasting products for people in different age and location groups. Thirdly, these city-level forecasts are distributed to people all around the country through a high viewership evening weather program on TV, allowing me to collect the actual forecast information broadcast to the general audience through transcribing these video recordings.

The day-ahead weather forecasts at city level are produced and distributed by the national forecasting system, managed under the CMA subordinate National Meteorological Centre (NMC)[3]. These forecasts are traditionally trusted partly because they are scientific forecasts generated by multiple well-established numerical models with inputs from high resolution atmospheric observations, and summarized by professional meteorologists[4]. For

[2]CMA is founded in 1949 and reformed in 1994 (Source: http://www.gov.cn/banshi/qy/rlzy/2012-11/12/content_2262675.htm). It is in charge of multiple duties in meteorology, including managing the national weather forecasting system as one of the main duties. The largest portion of CMA's annual budget is allocated to it, spendings include for maintaining and developing forecasting technologies, using new satellites, hiring and training professional forecasters. Example of 2012 CMA budget report: http://www.gov.cn/gzdt/att/att/site1/20120424/1c6f6506c7f81100fc9f1f.pdf.

[3]Source: http://www.cma.gov.cn/zfxxgk/gknr/jgyzn/jgsz/zsdw/202008/t20200813_4673126.html.

[4]The day-ahead national weather forecasts are generated as following. Each day, multiple numerical models are run with inputs (e.g., historical climate variables) collected from observatory stations and weather satellites, and they output forecasts variables including the near future (usually up to 2-3 days) temperatures, precipitations and pressures. I consulted the office of CMA, they did not specify which forecast models they used. But they confirmed there are multiple models taken

decades, these forecasts have been the almost single and uniform information source to the general public, quoted and distributed by most media channels including papers, TV and Internet[5]. They were reported to be popular among different groups of people all across the country as they adapt this information in deciding their daily activities such as labor, traveling and outdoor leisures[6].

One major challenge for data collection of short-range weather forecasts (for this paper and literatures) is that local agencies do not keep the historical data of city level forecasts[7]. This paper circumvents this problem by directly accessing the forecast information distributed to the people as a public good through TV. As one of the most important products, the 24-hour ahead city level weather forecasts are distributed over the high viewership evening *Weather Forecast* TV program right after the national news[8]. As a result, these TV broadcast weather forecasts are the actual information received by the vast and diverse

into considerations, and the list of models had been updated over the years where old models were dropped and new models were added. Some of the old models used are posted and shared under their database at https://data.cma.cn/data/cdcindex/cid/0b9164954813c573.html. Next, professional weather forecasters trained and employed at local weather stations evaluate these raw prediction outputs, and summarize the "average" forecasts with their best judgments. Thirdly, CMA holds the daily conference with local forecasters across the country, collecting their final local forecasts, and verifies their consistency at the national level (source: http://www.cma.gov.cn/2011xzt/2013zhuant/20130524/, https://zhuanlan.zhihu.com/p/21598589). Lastly, the meteorological center distributes and broadcasts the national forecasts to the general public through radios, TV, papers and Internet.

[5]Alternative sources of forecast information have become more common over the past decade, as new mobile apps like the Apple Weather generate their own weather forecasts. The national weather forecasts source still remain quite important to people around the country (source of recent survey conducted by CMA: https://mp.pdnews.cn/Pc/ArtInfoApi/article?id=26003677.

[6]Nationwide CMA survey in 2006 source Yuan, Sun and Wang, 2016, most recent survey conducted in 2021 https://mp.pdnews.cn/Pc/ArtInfoApi/article?id=26003677.

[7]I ask a CMA person and he quoted "lack of storage spaces" as the main reason.

[8]The *Weather Forecast* program was first aired in 1980, since then become the most popular weather TV program of the country (source:http://www.weather.com.cn/video/tqyb/05/508815.shtml. The program is aired daily on CCTV Channel 1 (China Central Television - Main Channel) and Channel 13 (News Channel) almost immediately after the national news concluding at 7:30pm. Until today, the evening *Weather Forecast* program is with the highest viewership among all TV programs year round (TV viewership reference data: https://eye.kuyun.com, http://www.csm-huan.com).

audience around the country[9]. Even if more people do not watch TV for weather forecasts over the most recent years, the same weather forecasts are distributed to alternative distribution channels they have accessed to from the dominant source of CMA. I then collect these day-ahead weather forecasts through directly transcribing the video recordings for over 2000 days of this *Weather Forecast* program. Specifically, I download a total over 130 hours of the program videos posted on the official site CCTV.com and transcribe all broadcast information feeding transformed audios to Google Cloud speech-to-text API. For detailed description of this data transcription process, see Appendix 1.A.1.

My video transcription dataset contains a panel data of the day-ahead forecast temperatures (daily temperature range including the minimum and maximum, $Tmin$ and $Tmax$) and category (e.g., sunny, shady, light rains, fog) over the next 24-hour, from 8pm today to 8pm tomorrow. As labor literatures like Graff Zivin and Niedell (2014), I focus on the daily maximum temperature forecasts for labor response, because labor activities are mostly during daytime when $Tmax$ is realized. Due to copy right issues accounted by NMC, my data starts from 2010 with missing days especially among early years (prior to 2013) where videos are either missing or corrupted from the website. Forecasts are issued at city level for each of the 34 provincial capital cities, centrally-administered municipalities and special administrative regions (SARs) across the country.

Historical weather forecasts for non-capital cities are more difficult to collect. Though

[9]According to CMA surveys, TV is an important channel for the general public for weather forecasts. In 2006, over 85% respondents receive weather forecast information through TV, dominant across all age groups (Yuan, Sun and Wang, 2016). The ratio drops to 22%-24% in the 2021 survey (source: https://m.thepaper.cn/baijiahao_16287375), but still remains the fourth most important information source.

each province has their local channels to broadcast their own weather forecasts program covering both capital and non-capital cities in the province right before the 7pm national news, the viewership is much smaller and these program videos are relatively scarce on the Internet. Due to time and resource constraints, I abstract away from these local forecasts and focus on possible extrapolations from national forecasts since viewership data indicates that people living in non-capital cities still take in the national forecasts seriously. As an alternative, I assume a simple mechanism that people living in non-capital cities still watch the *Weather Forecast* program (which is true according to viewership), and infer their cities' weather forecasts from the forecasts of their capital cities and historical weathers. Specifically, I approximate temperature forecasts for any non-capital city by adjusting the forecasts of its provincial capital with the difference in previous year monthly average realized temperatures (source ERA-Interim, see next subsection) between capital and non-capital cities[10]. For details about this approximation process and statistical tests regarding its credibility, see Appendix 1.A.2. Later analysis is also conducted on a sample restricted to capital cities only and main results remain robust.

I further restrict the forecast sample to Mainland (31 provinces) excluding the SARs where the *Weather Forecast* program is less popular among local audience. My final dataset covers 342 cities in total, 31 provincial capitals with direct forecasts, and 311 non-capital cities with approximated forecasts as addressed above. For this paper, I use the forecasts of six years 2010 to 2015, covering 2097 days with non-missing data.

[10]Ideally for ongoing data works, existing videos of local weather forecasts would be transcribed and compared with the collected capital city forecasts to derive relationships between non-capital city forecasts, capital city forecasts and historical realized weathers using machine-learning.

1.2.2 Realized Weather Data

Realized weather records are important to determine whether weather forecasts are accurate. In this paper, I use realized weather data from ERA-Interim (ERAI) reanalysis climate data product, produced by ECMWF (European Centre for Medium-Range Weather Forecast)[11]. The ERAI reanalysis data product uses mathematical models to extrapolate on existing station recordings of weathers, therefore not identical to the raw historical weather series from weather station readings only. Climate scientists normally judge this data product as sufficiently close to the "real" weathers (Dee etal., 2011), especially for temperatures, though there are more uncertainties around precipitations (Copernicus, 2017). As a result, the ERAI data product has been used in a variety of science and economic researchers as the realized weathers.

The ERAI dataset has the advantage of greater world coverage with higher spatial and temporal resolutions. It provides weather variables for all days post 1979 with the highest data frequency of 12 hours, reported on 0.25×0.25 grids all across the globe. To match the spatial and temporal resolutions of my forecast dataset, the ERAI data is aggregated to city-day level spatially by population weights[12] since in this paper the labor sample surveyed individuals who are likely residing and working in more populous grids. Alternative aggregation by area weights is also tested for robustness, and changes to main results of later sections are negligible. The final realized weather dataset I obtain includes daily surface

[11]Website: https://www.ecmwf.int/en/elibrary/8174-era-interim-archive-version-20. For future analysis, the most recent ERA5 data product will be considered (results are not expected to be very different since predictions of the two models in population dense regions like China would be close).

[12]Gridded population of the world for 2010, data use and aggregation process authorized by Climate Impact Lab, the Energy Policy Institute at the University of Chicago.

temperatures $Tmin$, $Tmax$, and daily total precipitations from 2009 to 2016.

1.2.3 Labor Data

For this paper, I take the labor decision of how many hours to work as a behavioral variable with likely response to weather forecasts on a daily basis. If people expect uncomfortable temperatures such as extreme heat, they will decrease their labor hours to avoid increased health risks. As discussed in related literatures (Graff Zivin and Niedell, 2014; Garg, Gibson and Sun, 2020; Rode etal., 2022), this makes labor decisions an important adaptation in response to temperature shocks.

To quantify labor decisions, I use the individual worker self-reported labor time-use data sourced from China Health and Nutrition Survey (CHNS), household and individual datasets[13] as my labor variable. The CHNS project is conduct once in 2-4 years, sampling with a multistage, random cluster process in 15 north-eastern, central, eastern and south-western provinces, accounting for 47% of the country's population and over half of the national GDP in 2010 (Zhang etal., 2014). To match with my forecast dataset, my labor sample covers 2 years of survey (2011 and 2015) across 52 cities from 12 provinces, with the survey period spanning 10 months in the second half of the two years (July-December 2011,

[13]This research uses data from China Health and Nutrition Survey (CHNS) (https://www.cpc.unc.edu/projects/china), grateful to research grant funding from the National Institute for Health (NIH), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) for R01 HD30880, National Institute on Aging (NIA) for R01 AG065357, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) for R01DK104371 and R01HL108427, the NIH Fogarty grant D43 TW009077 since 1989, and the China-Japan Friendship Hospital, Ministry of Health for support for CHNS 2009, Chinese National Human Genome Center at Shanghai since 2009, and Beijing Municipal Center for Disease Prevention and Control since 2011, and thanks to the National Institute for Nutrition and Health, China Center for Disease Control and Prevention, Beijing Municipal Center for Disease Control and Prevention, and the Chinese National Human Genome Center at Shanghai.

September-December 2015).

The labor take-up variable I use is the weekly labor supply in unit of hours, the answer to the CHNS individual level survey question asking what is the participant's total number of hours working in the previous week. I link each observation of labor with my forecast dataset by city of residence[14] and date of survey[15]. When merging with the forecast data, I keep only observations with non-missing $Tmax$ forecasts across all seven days during the labor reporting week[16]. Various demographic variables and household characteristics surveyed by CHNS are kept in the labor dataset for later heterogeneity analysis, including age, gender, education, previous year employment and income (2015 currency with inflation adjusted by city level CPI provided by CHNS). Focusing on the working population, I limit the survey sample to adults aged $16 - 65$.

My merged labor sample size includes $11,012$ individual level reports of non-missing weekly labor hours[17], almost evenly splitted between 2011 (5802) and 2015 (5210). Individuals reported in this sample are all employed in 12 occupation categories plus others or missing, covering workers of various types with both low and high incomes, in urban and rural areas. Since the sample does not contain the lower GDP and less population dense

[14]Spatial cross-work table from Volume I, p659-p662 of the published book "The Health and Nutrition Conditions of Residents in Eight Provinces of China" by Keyou Ge, 1998. The greatest level of spatial identification is district or county, but I take city level because people are likely to work in the same city but different districts as their residence. Robustness check is conducted with available data matched at district/county level, and the main results of this paper have almost no changes.

[15]To match the time scale of the labor variable, weather and forecast variables are also aggregated to weekly by the previous natural week of the survey date.

[16]The attrition rate is 3% with observations dropped in 15 cities. They only apply to 2011 where more forecast videos are missing.

[17]Some but not all individuals report in both surveys 2011 and 2015.

western provinces, results of this paper may not apply to these underrepresented regions of the country. But on a population weighted average, this sample has city characteristics representative of the national average in 2011 and 2015[18]. The dataset may have another problem with bunching, where self-reporting errors make relatively larger mass of weekly labor observations by multiples of 5 or 7, especially the 40 hours work per week (8 hours per day over 5 weekdays as the standardized working time). Further works with alternative Chinese labor dataset will be desirable for robustness.

1.2.4 Control Variables

I collect city-year or city level environmental and socio-economic variables from various data sources for control. Demographic and macroeconomic variables (for example, GDP per capita adjusted to 2015 value by GDP deflater source World Bank, population, city area) come from China City Statistical Yearbook, 2011-2019[19]. Complimentary geographical variables including administrative boundaries come from the 2017 release of State Bureau of Surveying and Mapping[20]. Elevation data is from the World Bank[21]. Number of weather stations is obtained by counting stations within each city boundary, latitudes and longitudes posted by CMA.

[18]Comparison of a selected set of climate and macroeconomic variables between the national and labor sample cities by population weight is presented in Appendix 1.H.1.

[19]the electronic tables are published on CNKI.net. The yearbook records city level data for the previous year of its publication. All these variables vary by city-year, except I take city areas and water resources as time-invariant from the most recent 2019 yearbook.

[20]Applied from National Catalogue Service for Geographic Information. Grid files are summarized by 2018 city and district administrative codes to obtain boundaries, river lengths and lake areas.

[21]Pixel level elevations at https://datacatalog.worldbank.org/search/dataset/0037910. City level elevation is summarized by averaging across pixel points falling within each city boundary.

1.2.5 Summary Statistics

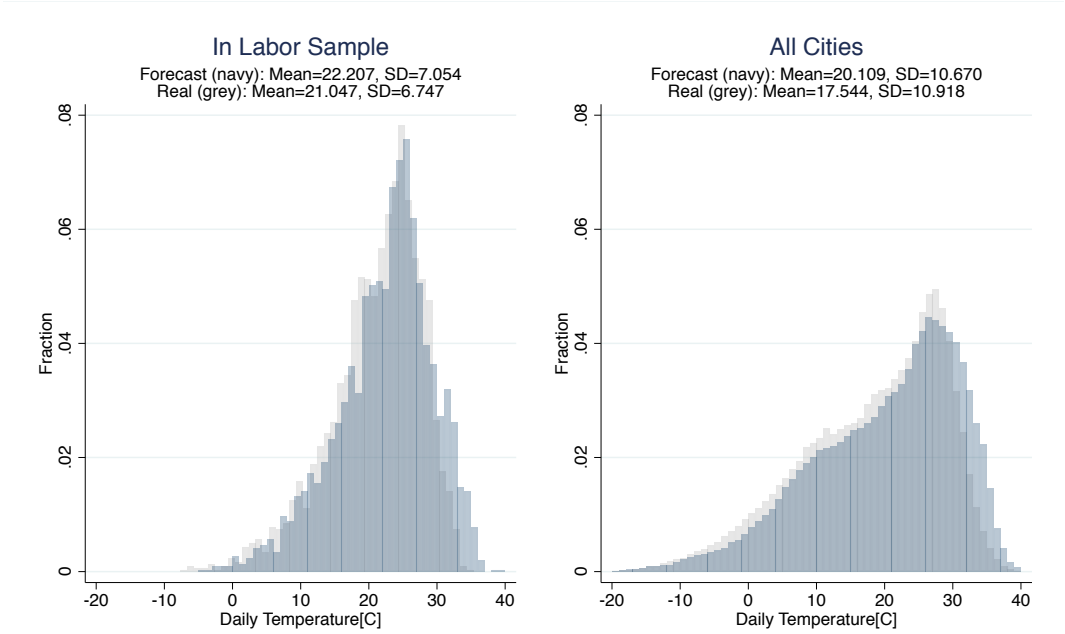
Figure 1.1 compares the city level daily forecast and realized $Tmax$ distributions. For the labor sample used in this paper (Panel (a)), realized and forecast temperature distributions are of similar shape (negatively skewed, with modes about $25C - 30C$, the usual comfortable daily maximum temperature range for human). Forecast distribution is to the right of the realized, by an average of $1.2C$. This bias implies negative forecast errors, or the daily weather forecasts of $Tmax$ usually overestimate the realized temperatures[22]. Higher moments are less different, with the standard deviation of the forecast distribution only $0.3C$ higher than the realized. The same observations apply for the larger sample containing all cities all days in 2011 and 2015 (Panel (b)). Because the full sample covers inland northern and western cities with more complex atmospheric dynamics that are harder to predict, its distributions have longer tails.

Summary statistics for the merged labor sample with $11,012$ observations are displayed in Table 1.1. The key labor variable, hours worked last week, is an average of 40.3 hours (5.76 hours averaged by 7 days or 8.06 hours averaged by 5 weekdays). Its range is relatively large, ranging from 0 to 24 (mostly service workers) hours worked every day of the week. Looking into weather variables over the survey weeks, daily $Tmax$ forecasts cover both hot and cold temperatures from $-5C$ to $40C$. Its mean is $22.3C$, again over-predicting by about $1.3C$ than the mean of realized $Tmax$[23]. Labor heterogeneity is featured in their demographic and the economic conditions. In the sample, 36.8% individuals are from provincial

[22]Actually, the percentage of city-days where $Tmax$ forecasts overestimate in the sample is 73.5%.

[23]Note these summary statistics are different from the previous histograms because they are weighted by number of individuals surveyed, while the histograms equal weight each city-day observation.

Figure 1.1: Daily Maximum Temperature Distribution, Forecast and Realized



Note: Forecast temperatures in navy, realized temperatures in grey. The left panel of *In Labor Sample* contains unique observations by city-day covered by the weeks reported in the labor surveys. The right panel of *All Cities* sample covers all 342 cities and all 730 days of 2011 and 2015 (685 days with non-missing forecasts). Histograms are capped by the minimum of $-20C$ and maximum of $40C$.

capital cities or centrally-administered municipalities and 40.6% from urban areas, leaving more than half that are likely in less developed areas. An average worker is about 43.8 years old, with more male (53.9%) than female (46.1%). Only 21.5% with reported education level has a college degree or higher, but the percentage is still greater than the national average[24]. From questions answered regarding previous year of the survey, workers worked an average of 10.3 months, earning monthly wages ranging from almost zero to above 1 million Yuan with a mean of $3,505$ Yuan. The sample represents a lower income group with annual wage income $37,219$ Yuan than the national average[25].

Table 1.1: Labor Sample Summary Statistics

Variable	N	Mean	STD	Min	Max
Avg. Hours Worked Last Week	11012	40.271	18.648	0.000	168.000
Daily Forecast Tmax [C]	77084	22.257	7.138	-5.000	40.000
Daily Real Tmax [C]	77084	21.004	6.804	-7.636	34.780
Weekly Precipitation [mm]	11012	17.018	21.049	0.000	109.898
In Capital City	11012	0.368	0.482	0.000	1.000
Urban	11012	0.406	0.491	0.000	1.000
Age	11012	43.752	10.987	16.000	65.000
Male	11012	0.539	0.498	0.000	1.000
College Degree or Higher	10672	0.215	0.411	0.000	1.000
Months Worked Last Year	10896	10.292	2.996	0.000	12.000
Monthly Wage Last Year [2015 Yuan]	6595	3505	20561	4	1149424
Wage Income Last Year [2015 Yuan]	7503	37219	59664	37	1967213

Note: The full labor sample includes $N = 11,012$ observations by individual-week-survey (2011 or 2015); Daily forecast and realized $Tmax$ are the only two variables summarized across all days included in the reported labor week; Wages are inflation adjusted to 2015 currency by city level CPI.

[24] In 2021 the number is about 15%, source National Bureau of Statistics of China.
[25] $53,615$ Yuan in 2015, source National Bureau of Statistics of China.

1.3 Empirical Strategy

1.3.1 Empirical Motivation

The empirical goal of this paper is to demonstrate there are labor response to forecast temperatures. The identification strategy would necessarily assume that decision makers take into consideration forecast information, and at least partially determine their labor choices based on forecasts ahead of weather realizations. This is suggested by Hsiang (2016), stating behavioral responses like labor can be dependent on both realization and belief about weathers. And weather forecasts are usually believed to be important but imperfect information in shaping the public's belief of future weathers.

The first empirical strategy is the simple regression estimating the average labor response to forecast temperatures. Similar to the models of labor-climate response from climate economics literature such as Graff Zivin and Neidell (2014), Garg, Gibson and Sun (2020), Rode etal. (2022), such relationship can also be addressed with the regression on realized temperatures. Results are expected to be very similar, because forecasts are highly correlated with realized temperatures[26]. Particularly, both simple regressions on forecast or realized temperatures are likely to find that labor hours decrease under uncomfortable temperatures like extreme heat, because of low productivity and increased health risks (Heal and Park, 2015).

The simple regression model misses an important fact, that labors may respond differently

[26]In my sample, the correlation is over 95%, and increases to over 97% if aggregate to weekly.

to accurate versus inaccurate forecasts. So instead for my main analysis, I will incorporate the impacts of average forecast accuracy perceived from historical forecast performance, which may be more important in shaping people's beliefs leading to labor decisions. In reality by behavioral economic theories, overreacting to not-that-extreme longer run average weather forecast errors is possible when people are making decisions under uncertainties, or knowing the potentials to endure losses (Twersky and Kahneman, 1974; Kahneman, Knetsch, and Thaler, 1991).

My baseline regression is then the simple regression adding the interaction with a second treatment representing the perceived forecast accuracy in medium-run (defined in next subsection). In this setting, my hypothesis is that individual decision makers plan their labor activities according to weather forecasts as well as the perceived medium-run average forecast accuracy over the past. Therefore, this regression estimates the labor response to forecast accuracy. By this, I also solve the problem of indistinguishable effects from realized temperatures with only the simple regression[27].

In an ideal setting, both weather forecasts and forecast accuracy will be randomly assigned to individual decision makers, which is difficult to obtain even with field experiments[28]. In this paper, I will argue that both forecasts and forecast accuracy are probably exogenous as largely determined by the same numerical modeling uniformly applied to all regions of the

[27]Because of the high correlation between forecast and realized temperatures in my sample, controlling realized temperatures directly in the simple regression will cause biased estimates due to multicollinearity.

[28]In that case, I would need to analyze the changes in labor outcomes before and after individual decision makers are randomly provided with no, low and high quality weather forecasts. But since people in the experiment can easily receive outside information about near-term forecasts provided as a public good, even this RCT setting will be hard to achieve.

country, and that meteorologists engaged in the forecasting process are roughly rational in making their best judgments about the model outputs. To test this assumption, I will also control interactions with observable economic and physical factors that may have impacts on both forecasts and labor for robustness check against omitted variable bias (OVB).

My research design is different from other literatures on short-range weather forecasts, in terms that I argue the labor reaction to forecasts through observing differential labor responses under different perceived forecast accuracy over historical forecasts. Lemoine (2018) proposes a model and empirical approach of regressions controlling both forecasts and realized weathers. Shrader, Bakkensen, and Lemoine (2022) uses an empirical design to address the outcome variable (mortality) respond to instantaneous forecast errors while controlling realized temperatures. These existing models focus on the impacts of instantaneous forecast errors instead, assuming that decision makers rationally access the forecast information with credibility.

1.3.2 Forecast Accuracy Metric

To complete the empirical design, I first need to define a forecast accuracy metric perceived by individual workers in my labor sample. When it is almost impossible to keep instantaneous forecast errors at almost zero everyday because idiosyncratic shocks in weather conditions are unpredictable by even the most advanced meteorological models, perceived accuracy of forecasts is more likely represented by the longer run average of forecast errors.

Consider i is the city and t is the date, $Tmax_{it}^{forecast}$ is the weather forecast broadcast at

the night of date $t-1$ predicting the maximum temperature of the next day t. I represent the perceived accuracy of this day t forecast with the root-mean-squared-error ($RMSE$) between forecast $Tmax^{forecast}$ and realized temperatures $Tmax^{real}$, over a medium-run length rolling window of the previous half-year:

$$RMSE_{it}^{Tmax} = \sqrt{\frac{1}{R}\sum_{s=1}^{R}(Tmax_{it-s}^{real} - Tmax_{it-s}^{forecast})^2}$$

The forecast accuracy metric $RMSE_{it}$ is defined as the square root of average sum of square errors $e_{is} = Tmax_{is}^{real} - Tmax_{is}^{forecast}$, realized middle of the day s, over all days s on the rolling window $t-1, t-2, ..., t-R$ with $R = 183$. As a result, my $RMSE$ metric summarizes the average forecast errors for nearest 183 days[29]. The lower the $RMSE$, the smaller the average forecast errors, and the more accurate the weather forecasts are perceived.

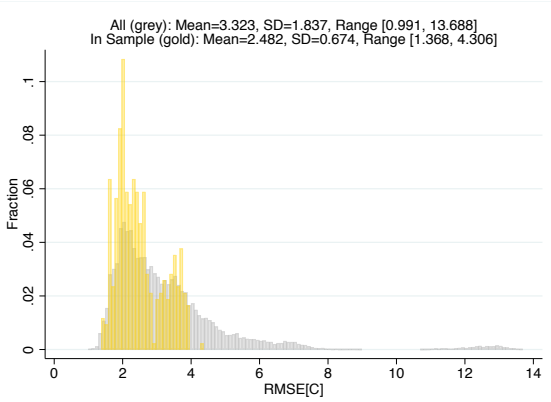
By definition, the $RMSE$ metric weights both bias and precision of forecasts (for analysis of the two breakdown quantities, see Appendix 1.H.2). For my main analysis, I choose a rolling window of $R = 183$ days (half a year) as it maximizes the explanatory power of the main regression (see Appendix 1.B). This rolling window $RMSE$ is not the unique metric of perceived forecast accuracy. Alternative metrics are tested later for robustness checks, and results do not diverge from the main specification and do no improve on the explanatory power of the baseline regression.

The resulting $RMSE$ metric varies both in time and in space. As summarized in Figure

[29]Forecast error for $t-1$ is realized midday of date $t-1$, hours before the date t forecast is aired on TV.

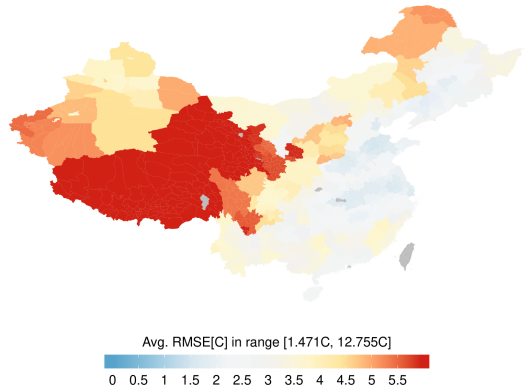
Figure 1.2: Summary Plots for the Spatial and Temporal Variations of Forecast Accuracy Metric, Half-Year Rolling RMSE.

(a) The Distribution of $Tmax$ Forecast RMSE on Half-Year Rolling Window



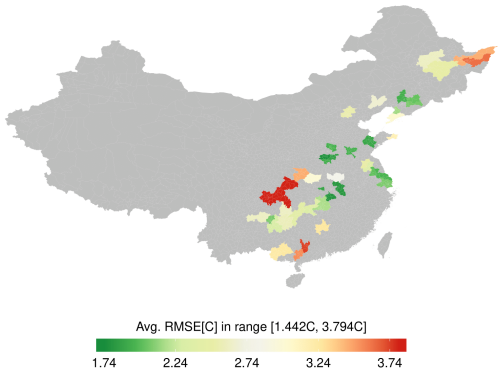
Note: Distribution is for half-year (183 days) rolling $RMSE$ by city-day. Gold is only the labor sample coverage (N = 424, on weeks covered by the labor report, $RMSE$ of each week is taken as the value perceived on Monday), Grey is all 342 cities of all days 2011 and 2015 with non-missing forecasts (685 days, total N = 234270.

(b) The Spatial Distribution of Average Half-Year Rolling RMSE By Cities, All Cities



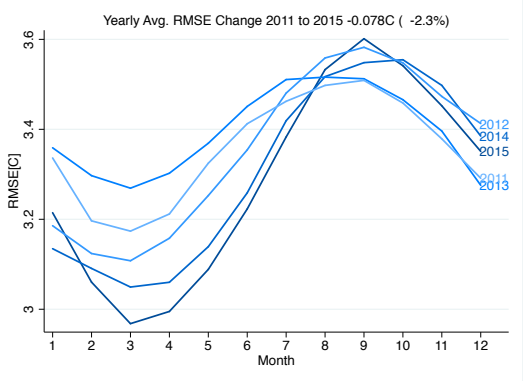
Note: The map illustrates the spatial distribution of half-year rolling forecast $RMSE$ of daily $Tmax$. The city-day observations of rolling $RMSE$ are averaged by all 342 cities over 685 days with non-missing forecasts in 2011 and 2015.

(c) The Spatial Distribution of Average Half-Year Rolling RMSE by Cities, Labor Sample Only



Note: The half-year rolling RMSE is taken at Mondays of the weeks with labor reports. N = 424 city-day observations covered by the labor sample. Map displays the spatial distribution of this sample average $RMSE$ by 52 cities.

(d) The Trend of Average Half-Year Rolling RMSE By Months



Note: Rolling $RMSE$ of all 342 cities and all days with non-missing forecasts are taken for 2011-2015 and averaged by months. Plot summarizes the 12-month trend of this average $RMSE$, for each of the 5 years in sample. Lighter color lines indicate earlier years.

1.2, Panel (a) shows that $RMSE$ spans a range of different values from about $1C$ to above $4C$, featuring significant mass around the mode $[2C, 3C]$. In Panel (b), spatial variations are shown to be quite significant and persistent both within and across province boundaries as well as topographical features (e.g., rivers and mountains). Very large $RMSE$ occurs for northern and western inland cities, where weather predictions are harder with higher elevations and less stable climate conditions, but of small population and not included in the labor sample of Panel (c). The labor sample average still covers a visible range of $2.35C$, with higher $RMSE$ cities to the north-east and south-west of the country.

Meanwhile, temporal variations in $RMSE$ are much smaller in magnitudes because of smoothing over the half-year window. In my labor sample, 82.8% of total variations are contributed by spatial (between) variations. As indicated in Panel (d), there is a seasonal pattern when $RMSE$ is maximum around fall and minimum around spring (resulting from smaller forecast errors over colder months), but the size of their average difference is only about $0.6C$. Total $RMSE$ averaged across all cities and dates of the year only decreases by $0.078C$ (2.3%) from 2011 to 2015. These decreases are non-monotonic, mostly coming from the spring months.

An important question related to the exogeneity assumption of empirical design is where these variations in forecast accuracy, especially the spatial variations, come from. Based on the procedure of Chinese weather forecasting, forecast inaccuracy can arise from various factors. Among the observables, economic and physical conditions are by design likely related to forecasts performance. To examine that, I run the simple OLS regression of $RMSE$ on

the sets of selected economic and physical factors in Table 1.C.1 (economic factors) and 1.C.2 (physical factors)[30]. Both economic and physical factors explain some variations in $RMSE$, with statistically significant coefficients for city area[31], population, share of labor force, annual mean and standard deviation of realized $Tmax$[32], number of weather stations, greenland area and length of rivers. Physical factors contribute more than economic factors[33], but neither can explain all the variations of $RMSE$[34]. The left over variations are can related unobserved factors from more exogenous events[35].

1.3.3 Regression Design

As discussed in previous subsection, I first run a simple regression estimating the average labor response to forecast temperatures. This regression design is similar to previous literatures, such as Graff Zivin and Neidell (2014), but replacing the response to realized with forecast temperatures:

[30]The regression is run on distinct observations of $RMSE$ by city-day covered by my labor sample. Observations with missing economic or physical factors are dropped, remaining N = 9507. These city level control variables are either invariant in time or averaged by years, therefore mainly representing the spatial variations. To capture temporal variations also, I allow a date fixed effect δ_t in column (3) of these tables. Since the labor sample is taken near end of the year, current year average may be more representative for the climate condition. An alternative with previous 5-year average temperature and precipitation is also run, giving similar results. The same regressions on the extended sample of all cities and all days with non-missing forecasts in 2011 and 2015 are summarized in Table 1.C.3 and 1.C.4 with similar conclusions.

[31]Positive significant estimates consistent across all three columns, meeting the intuition that larger cities are harder to cover with forecasts therefore come with greater $RMSE$.

[32]The negative coefficients are more counterintuitive, implying that hotter cities with more daily maximum temperature variations enjoys lower $RMSE$, while the general belief is that more extreme and more variant weathers are harder to predict.

[33]Comparing R-squared of columns (1) of the two tables, physical factors has greater explanatory power than economic factors (0.489 vs 0.405), consistent with the theory that the physical conditions are more correlated because they enter the scientific process generating those forecasts.

[34]The full regression column (2) containing both sets of factors explains just below 60% of the data, leaving over 40% unexplained variations in the residual term. As expected in column (3), further controlling for time fixed effect does not have great impacts on the regression results, still leaving over 30% to unobserved factors.

[35]For example, the limitation of numerical models, the introduction of new meteorological satellites, the decrease of human mistakes due to improved trainings, and the influence of global climate events like El Nino.

$$Labor_{ikt} = f(Tmax_{it}^{forecast}; \beta) + \gamma' \mathbf{X}_{it} + \epsilon_{ikt} \tag{1.1}$$

Labor response to next-day forecasts is estimated as a non-linear function of $Tmax^{forecast}$. Note that index k is per person surveyed, i is the city where the person resides, t is the day that forecasts target. The control set $\mathbf{X}$ includes a quadratic function of real precipitations (i.e., controlling for precipitation and its square). Month and city fixed effects are also included. All variables on the right hand side (except for fixed effects) are summed by the same natural weeks[36] as the outcome variable of weekly labor hours to match the data frequency. Standard errors are clustered at city level.

The baseline regression design is built on the simple regression Eq. 1.1 with the interaction of half-year rolling $RMSE$. The $RMSE$ metric represents the perceived forecast accuracy, entering the regression independently as the linear control as well in $\mathbf{X}$. This adapts an empirical design similar to Carleton etal. (2020), estimating the differential labor responses to forecast temperatures under space-time varying perceived forecast accuracy measured by the covariate $RMSE$:

$$Labor_{ikt} = f(Tmax_{it}^{forecast}; \beta_0) + f(Tmax_{it}^{forecast}; \beta_1) \times RMSE_{it} + \gamma' \mathbf{X}_{it} + \epsilon_{ikt} \tag{1.2}$$

Here I take the $RMSE$ metric to be uniform across the labor reporting week as realized on the first day (Monday), since the temporal variation of $RMSE$ across seven days is

[36] Monday to Sunday.

small[37]. All settings remain the same as the simple regression, with $\mathbf{X}$ containing precipitation and its square, city and month fixed effects besides $RMSE$ as a linear control. In robustness checks, $\mathbf{X}$ with more linear controls and more stringent fixed effects are tested. Realized temperatures are not controlled in the baseline regression due to multicollinearity, but will be included for robustness. Standard errors are again clustered at city level.

For main analysis, the non-linear labor response function $f(.)$ is taken to be a restricted cubic spline with 5 knots, $(5C, 15C, 20C, 25C, 35C)$. Appendix 1.B discusses the specification choice and justify the positions of knots based on balancing explanatory power and precision of regression estimates for both Eq. 1.1 and Eq. 1.2). Robustness checks later will perform analysis on alternative functional forms (non-parametric $5C$ bins and 3 knots linear spline). The function parameter vectors β (β_0, β_1 in Eq. 1.2) are the key coefficients for estimation, with which labor response to forecasts and forecast accuracy are identified by these regression models.

To analyze the heterogeneity in labor response estimated by Eq. 1.2 by labors' geographical, macroeconomic and demographic characters, I will also run the following regression separating my sample by groups $g \in G$, with the set of group indicators $\mathbf{1}_g$ also included as linear controls:

[37]Robustness is run for the explicit version allowing $RMSE$ to vary across days of the week, but main results have almost no changes.

$$Labor_{ikt} = \sum_{g} \mathbf{1}_g [f(T_{it}^{forecast}; \beta_{0g}) + f(T_{it}^{forecast}; \beta_{1g}) \times RMSE_{it}] + \gamma' \mathbf{X}_{it} + \epsilon_{ikt} \tag{1.3}$$

Lastly, as argued in previous subsection, I run Eq. 1.2 controlling additional interaction terms with potential omitted variables (OVB) possibly correlated with $RMSE$ and are also determinants for the labor-forecast responses to address robustness of my baseline results:

$$\begin{aligned} Labor_{ikt} = {} & f(Tmax_{it}^{forecast}; \beta_0) + f(Tmax_{it}^{forecast}; \beta_1) \times RMSE_{it} \\ & + \sum_{m \in M} f(Tmax_{it}^{forecast}; \beta_m) \times m_{it} + \gamma' \mathbf{X}_{it} + \epsilon_{ikt} \end{aligned} \tag{1.4}$$

Here M is the set of potential omitted variables with $m \in M$ as additional linear covariates besides $RMSE$. In later robustness checks, I first consider m to be the average income and climate variables commonly seen in climate literatures. I also use the saturated set with all physical and economic factors possibly correlated with $RMSE$ as discussed in previous subsection. Each m is also included linearly in the control set $\mathbf{X}$.

1.4 Results

1.4.1 Main Results

Main regression results are presented in Figure 1.3. The simple regression (Eq. 1.1) is in Panel (b), and the baseline regression with $RMSE$ interaction (Eq. 1.2) is in Panel (c) and

(d). All three panels of labor responses are plotted against the horizontal axis of forecast daily maximum temperatures in the range $0C - 40C$[38]. These labor responses are re-centered at a reference temperature of $Tmax^{forecast} = 25C$, approximately the daily maximum temperature where human feels comfortable (Graff Zivin and Neidell, 2014; Carleton etal., 2020; Rode etal., 2022). These plots represent the changes in daily labor hours under the forecasts $Tmax^{forecast}$, relative to the reference of $25C$[39].

First from the simple regression (Eq. 1.1) results in Panel (b), daily labor hours decrease under high daily forecast temperatures relative to the reference of $25C$. But the magnitude of such decrease is small, about $13 - 16$ minutes, and statistically insignificant at 5%. As expected giving the high correlation between forecast and realized temperatures, this regression result is similar to the labor response estimated with realized temperatures (Figure 1.D.1) and cannot be interpreted directly as labor decisions reacting to forecasts. The magnitudes of these hot end estimates are consistent with previous literatures which give estimates of $17 - 22$ minutes decrease of labor per day under extreme heat ($40C$) (Garg, Gibson and Sun, 2020; Graff Zivin and Neidell, 2014; Rode etal., 2022). Unlike these literatures showing either flat or negative labor decreases under the cold (below $25C$), my global regression results have an uncommon jump of labor up to half an hour (30 minutes) under the medium-cold

[38]This range accounts for 99.8% of the sample temperatures, dropping only $Tmax^{forecast} < 0$. I focus more on the hot end responses where labor take-up decreases due to health concerns are more evident in previous literatures.

[39]The main reason to recenter at $Tmax^{forecast} = 25C$ is because by design any labor response estimated directly from the regressions would be relative to $0C$, which is a cold temperature hard to interpret for intuitions. But under a comfortable temperature forecast of $Tmax^{forecast} = 25C$, I can expect maximum labor supply choices. I can further assume little labor variations under this reference forecast with changing $RMSE$ in the sample range $[1.4C, 4.3C]$, as the realized temperatures are likely still comfortable enough for labors to choose their maximum working time.

forecast range of $15C$ – $20C$[40]. One explanation can be that there are overlooked factors shifting the labor-temperature response such as forecast accuracy[41].

From the baseline regression (Eq. 1.2), I estimate the labor responses to forecast temperatures from $0C$ – $40C$ under different perceived forecast accuracy scenarios $RMSE$ = $3C, 2C, 1C$ in Panel (c)[42]. From left to right, when forecasts are inaccurate at $RMSE = 3C$, I observe an almost flat and statistically insignificant labor response across all forecast temperatures relative to $25C$, except for the positive labor change around $20C$. Then moving towards more accurate forecasts when $RMSE$ decreases towards $1C$, I observe the large and 5% statistically significant labor decrease relative to $25C$ over two ranges of forecast temperatures. One is the hot temperatures $Tmax^{forecast} \geq 30C$, and the other is the medium-cold temperatures $15C \leq Tmax^{forecast} < 25C$. On the other values of forecast temperatures including the cold end $Tmax^{forecast} < 15C$, labor response remains statistically insignificantly different from the reference $25C$ regardless of $RMSE$. Therefore, there are large and signif-

[40]This observation is actually quite consistent across different functional form choice $f(.)$, or with the estimation under realized temperatures Figure 1.D.1.

[41]It can also because this paper uses a different sample range and including more recent years of the CHNS surveys compared with the previous study of Chinese labors (Garg, Gibson and Sun, 2020).

[42]$RMSE$ range is based on in-sample distribution. $RMSE = 4C$ is not selected because only 0.15% observations have $RMSE \geq 4C$. The sample even extending to all cities and days 2011 and 2015 features $RMSE$ with almost no weight below $1C$, making labor responses under lower $RMSE$ towards the "perfect" forecasts at $RMSE = 0C$ extrapolations (in real life, perfect forecasts are impossible because climate conditions change with stochasticity and randomness). In Appendix Figure 1.D.2 I still illustrate the $RMSE = 0$ case extrapolated by my baseline regression model, featuring very large and statistically significant labor decrease under extreme hot and medium-cold. The decrease of labor under perfect forecasts at $40C$ is very large enough to reduce daily labors to zero. For more realistic representations, I also select five large provincial capital cities (Jinan, Changchun, Beijing, Kunming and Chengdu) with increasing average $RMSE$ from about $1.5C$ – $3.8C$ over all days with non-missing forecasts 2011 and 2015. Under their different levels of average $RMSE$, the labor response is estimated for each city using Eq. 1.2 in Table 1.D.1. Conclusions are similar to Panel (c), showing negative labor responses relative to $25C$ under hot ($Tmax^{forecast} = 35C, 40C$) and medium-cold ($Tmax^{forecast} = 20C$) temperature forecasts for cities with smaller $RMSE$ only (Jinan and Changchun), but not significant and even the reverse positive for cities with higher $RMSE$ (Beijing, Kunming and Chengdu).

icant labor decreases under hot or medium-cold forecasts only when forecasts are perceived as accurate enough. Instead when forecasts are inaccurate, these labor responses diminish.

The same conclusion can be drawn from the $RMSE$ marginal effects graphed in Panel (d). For these two forecast temperatures ranges (hot and medium-cold), there are large magnitudes of 5% statistically significant positive marginal effects of the $RMSE$ covariate, while the other values of $Tmax^{forecast}$ correspond to near-zero and statistically insignificant marginal effects relative to $25C$. In interpretation, marginal decrease of $RMSE$ will lead to large and statistically significant labor reduction under hot or medium-cold temperature forecasts, contributing to the negative labor responses under these two forecast ranges under $RMSE = 1$ seen in Panel (c).

Focusing on the hot end response above $30C$, the marginal effect of $RMSE$ is monotonically increasing in $Tmax^{forecast}$, resulting in greater decrease of labor under hotter temperature forecasts when $RMSE \leq 2C$. At maximum, 2.4 hours of work are reduced under $Tmax^{forecast} = 40C$ per $1C$ decrease of $RMSE$, leading to the large relative labor reduction by 4.5 hours per day under $Tmax^{forecast} = 40C$ and $RMSE = 1C$. Even for not that extreme hotness under $Tmax^{forecast} = 35C$, the labor reduction perceiving $RMSE = 1C$ forecasts is as large as 2.3 hours per day, suggesting 28.9% decrease in labor supply (compared with standardized 8 hours of working per day)[43]. These values are not only large impacts of forecasts and forecast accuracy in reality, but also factor of tens greater than

[43]Considering the country's public policy ordering to pause outdoor labors during extreme heat (above 35C), these results can reflect a combination of the optimal labor choice on individual (intensive) and society (extensive) margins under expected heat waves. In the main section, I discuss as if it is a rational public policy consistent with labors' utility maximization.

the simple regression estimates in Panel (a) as well as previous literatures. However, these large estimated labor responses are likely not contributing significantly to the overall welfare changes because extreme hot forecasts represent only a small share of observations in reality ($Tmax^{forecast} \geq 35C$ accounts for only 1.2% in Panel (a)).

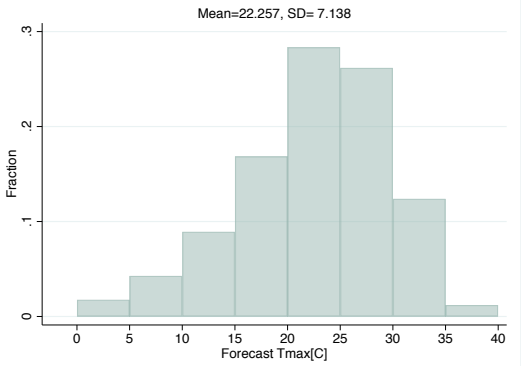
Labor response to medium-cold forecasts $[15C, 25C)$ is non-monotonic, smaller in magnitude but quite robust and statistically significant. When there are accurate forecasts with $RMSE = 1C$, people work up to 1.2 hours less under colder temperature forecasts in this range relative to $25C$. When there are medium forecasts with $RMSE = 2C$, people's labor hours are similar to under $25C$. When there are inaccurate forecasts with $RMSE = 3C$, people work up to 0.8 hours more. The maximum marginal effect and labor decrease under accurate forecasts happen around $Tmax^{forecast} = 20C$, with these effects decreasing both above and below this temperature. These responses are again large in reality, especially making greater welfare impacts with much higher frequency of occurrence for the medium-cold temperature cohort compared with extreme heat (45.2% of the sample from Panel (a)).

According to this empirical analysis, it has been suggested that labors respond to both forecasts and forecast accuracy. Related to theories in climate economics, main results can be explained by workers reacting to forecasts for the trade-off between wage income and utility gains from not working under uncomfortable temperatures. When people access the forecast information when forecasts are accurate, they are more responsive in labor decisions for health concerns. Otherwise if forecasts are perceived as inaccurate, they are reluctant to change their labor time at the costs of wage and consumption. For the hot end, this theory

matches the documented disutilities of labors working under heat. For the medium-cold range, similar argument applies but it is less seen that the labor decrease disappears when forecasts drop below $15C$[44]. A reasonable explanation would be that heating is turned on as temperature decreases, making labors less sensitive to forecasts[45]. In the next subsection, this hypothesis will be tested with a heterogeneity analysis by availability of heating.

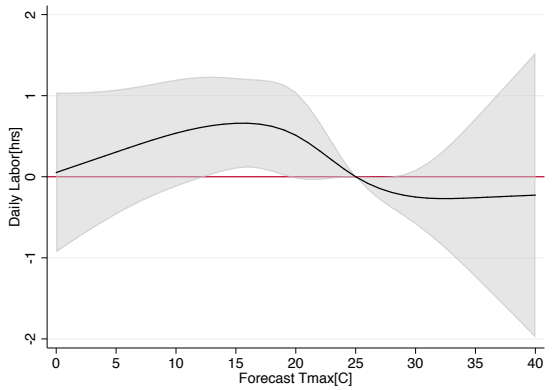
Figure 1.3: Main Regression Results - Forecast Temperature Histogram, Simple Regression, Baseline Regression with RMSE Interaction Array and Marginal Effect

(a) Daily Maximum Temperature Forecasts Distribution



Note: The histogram shows distribution of $N = 77,084$ observations of $Tmax^{forecast}$ covered by the labor sample (by individual-city-day, weighted by number of individuals surveyed as in Table 1.1). Observations are sorted into $5C$ bins with the end bins $(-\infty, 5C)$ and $[35C, \infty)$.

(b) Simple Regression



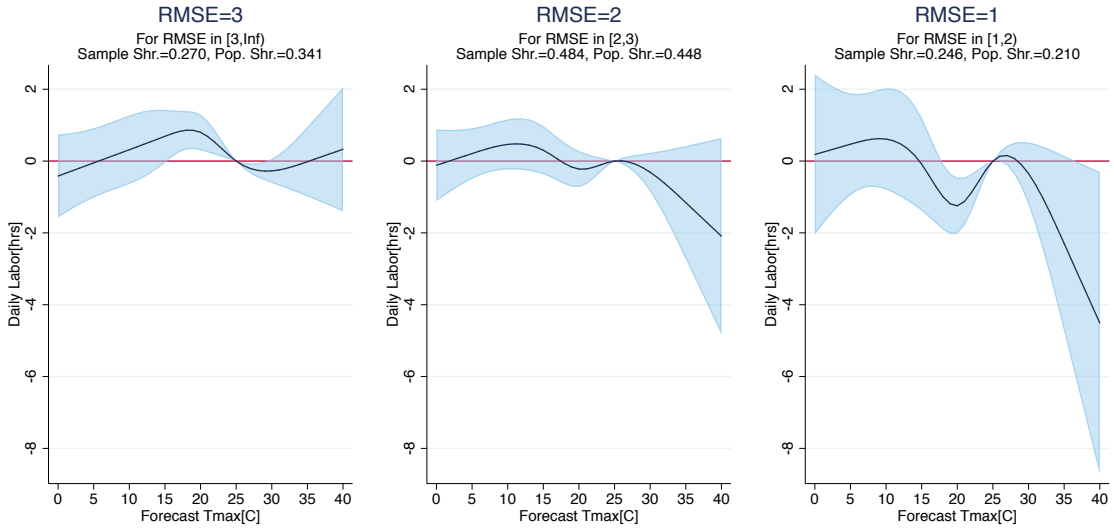
Note: The simple regression run is Eq. 1.1. Daily labor response is evaluated with the estimated $f(.;\beta)$ from the regression (keeping all other controls in $\mathbf{X}$ constant), and subtracting the reference labor response at $25C$. In interpretation, this plot indicates the labor response to forecast $Tmax$ relative to $25C$. The shaded area is the 95% confidence interval for the labor response.

[44]Other papers mostly focus on the hot end labor responses and do not find such labor response. At the cold end, some find little to no response (Graff Zivin and Neidell, 2014; Rode etal., 2022), while others find almost symmetrical labor decrease as the hot end (Garg, Gibson and Sun, 2020).

[45]According to WHO's suggestion (source: https://apps.who.int/iris/bitstream/handle/10665/275839/WHO-CED-PHE-18.03-eng.pdf) and the country's engineering standard for industrial and commercial working conditions of heating (GB/T18883-2002, GB 50019-2015, GB50019-2003), indoor temperatures in China are suggested to maintain at $16-24$ Celsius. This would indeed define below $15C$, especially for daily maximum temperature, as the range where heating should be provided in workplaces.

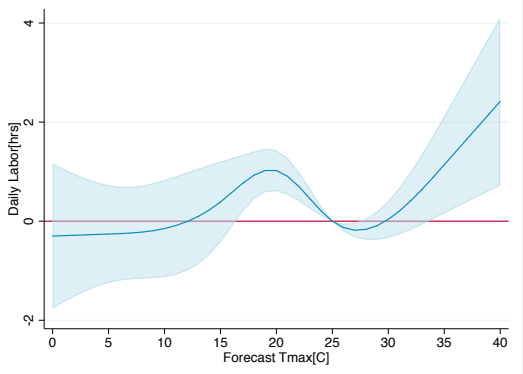
Figure 1.3, continued

(c) Baseline Regression with RMSE Interaction - Array of Estimated Labor Responses Under RMSE=3C,2C,1C



Note: Baseline regression 1.2 is run. Daily labor response relative to 25C is evaluated with the regression estimated functions $f(.;\beta_0)$ and $f(.;\beta_1)$ interacted with $RMSE = 3, 2, 1$, keeping other controls constant and subtracting the labor response at the reference temperature of $25C$. Shaded area represents the 95% confidence interval. From left to right, array displays the scenarios when $RMSE$ decreases from $3C$ to $1C$. The above-graph notes summarize the sample share of each $1C$ range of $RMSE$ by number of observations in the labor sample ($N = 11,012$), and the population share of 2015 by city average RMSE across the full 2011 and 2015 sample of 685 days and 291 cities with non-missing 2015 population (source: China City Statistical Yearbook).

(d) Baseline Regression with RMSE Interaction - Marginal Effect of RMSE



Note: The marginal effect of the linear covariate $RMSE$ is estimated by $f(.;\beta_1)$ from the baseline regression 1.2, relative to $25C$ keeping other controls constant. In interpretation, the curve represents the estimated labor change relative to $25C$ when there is $1C$ marginal increase of $RMSE$. Shaded area represents the 95% confidence interval.

1.4.2 Heterogeneity Analysis

An important question remained to answer is what types of labor in specific contribute to the large and significant labor responses to weather forecasts estimated in this paper. For this purpose, I run heterogeneity analysis separating my labor sample with Eq. 1.3.

Firstly, I test the hypothesis about heating availability and medium-cold labor responses. To do so, I run heterogeneity analysis separating the sample by northern and southern cities in China. These cities are traditionally defined by the Qin-Huai Line, where northern cities have greater accessibility to heatings partially due to their central heating infrastructures built by the government. Southern cities have individuals, households and corporates fully in charge of their heating needs, with more costly and less convenient air conditioners or smaller portable heating devices. By hypothesis, northern cities are more likely to have the medium-cold temperatures $15C - 25C$ covered under some preventive measures like heating, making labors less responsive to forecasts and forecast accuracy over this range, especially for those working indoors. Meanwhile, southern cities would be more sensitive to forecasts, until it is cold enough and heating is vastly turned on.

My results from regression Eq. 1.3 present the cold end arrays (hot ends are quite noisy) for northern and southern cities in Figure 1.4. The labor decrease in response to medium-cold forecasts under low $RMSE$ is seen only in southern but not northern cities, then affirming my hypothesis that heating may help the northern city labors to be less sensitive to work under medium-cold forecasts. There is also an alternative test in Figure 1.D.3 instead sub-

sampling for cold, moderate and hot cities by current year average realized $Tmax$ tercile groups[46]. Again, the labor decrease under medium-cold forecasts and low $RMSE$ only exists for hot and moderate terciles, but not the cold tercile.

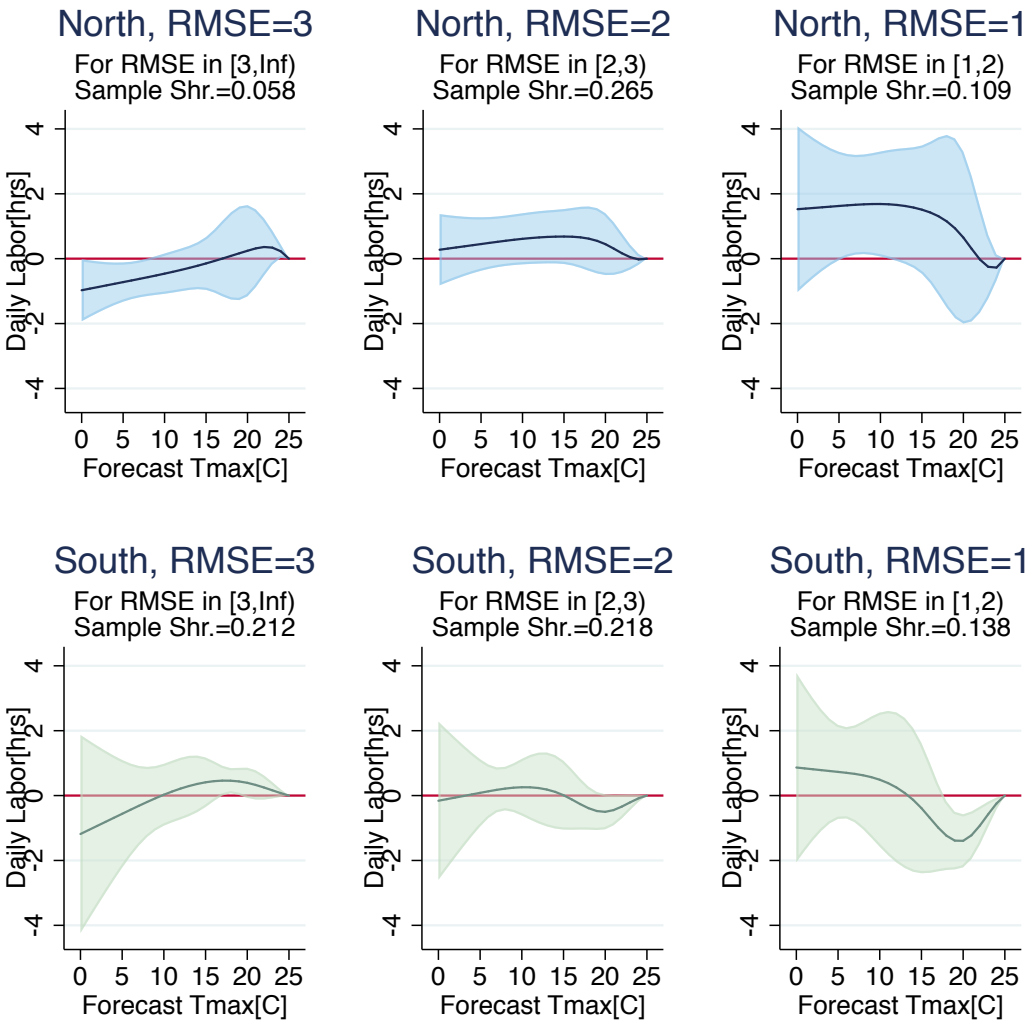
Secondly, I want to explore whether different groups of labors respond differently to forecasts and forecast accuracy. I first run the heterogeneity analysis with the CHNS primary occupation categories. Figure 1.D.4, 1.D.5 and 1.D.6 show the results for every occupation[47]. All categories have noisy estimates for both arrays and marginal effects, making any arguments not conclusive due to lack of statistical significance. However, I can still see typical outdoor occupations like drivers and farmers displaying results similar to the baseline, showing labor reduction under hot and medium-cold forecasts as $RMSE$ decreases. Instead for some typical indoor occupations like managers, or jobs with less flexibility such as policemen, the labor responses are either flat or opposite to the main results. Some categories like office staffs and entertainment business are probably indoor but still display some negative response to hot forecasts. Other occupations, such as service workers and technicians, are less apparent in their exposure to climate risks. Overall, these results are about consistent with previous literatures suggesting labor response to temperatures is mainly contributed by high risk workers of more climate exposures (Graff Zivin and Neidell, 2014; Rode etal., 2022).

For evidence with greater statistical power, I run the subsample analysis on two subgroups separated by 13 macroeconomic and demographic variables at city, household and individual

[46]Similar results are found using the previous five-year average or current year $20C$ based heating degree days of $Tmax$.

[47]Missing or non-specified are included in the regression but results are not presented here. That contributes to 5.9% of the sample.

Figure 1.4: Heterogeneity Analysis with North-South Separation



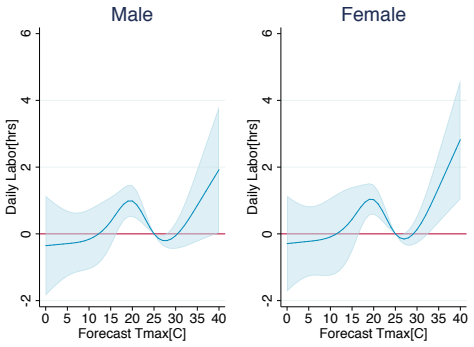
Note: Northern and southern cities are defined by the traditional separation with Qin-Huai line. Regression Eq. 1.3 allows different direct and $RMSE$ interacted marginal labor responses to forecast temperatures for northern and southern cities. The figure displays array for cold end ($Tmax^{forecast} \leq 25C$) responses relative to $25C$. From left to right, array displays the scenarios when $RMSE$ decreases from $3C$ to $1C$. Labor response of northern cities on top row, southern cities on bottom row.

levels, and present the $RMSE$ marginal effects comparison for each set in Figure 1.5. There are some indications that labor decrease under accurate uncomfortable temperature forecasts occurs only for more vulnerable labor groups (in health or in economic conditions) that can also be more sensitive to climate risks. Those include the lower degree holders (Panel (b)), the older population (Panel (c)), the lower household or individual incomes (Panel (d) and (e)), labors working fewer hours and earning lower hourly wages last year (Panel (g) and (h)), and workers living in rural districts (Panel (j)). These results match the classical labor theory that those gaining less from labor or lose more to climate shocks would have greater incentive to work less in trade-off with lower risks working under uncomfortable weathers they believe to come.

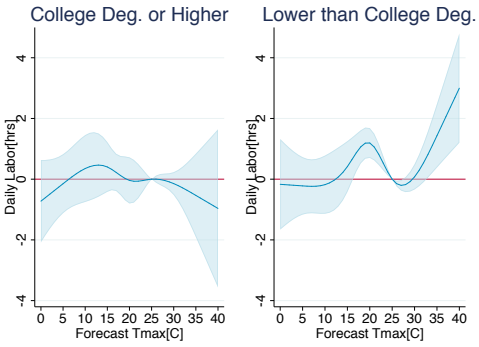
1.4.3 Robustness to Factors that Affect Labor-Forecasts Responses

In my baseline regression Eq. 1.2, omitted variable bias (OVB) may affect the key estimates of labor-forecasts responses. For robustness against OVB, I run Eq. 1.4 controlling sets of potential omitted variables both as additional interaction terms and separately as linear controls. Such variables are omitted variables if they are both determinants of the labor response to temperature forecasts, and correlated with the half-year rolling forecast $RMSE$ covariate. As argued before, the $RMSE$ metric can be correlated with the economic and climate conditions of each city. In previous climate economic literatures (Carleton etal., 2020; Hsiang, 2016; Rode etal., 2022), both conditions are important adaptation factors under which labor responses to temperatures vary. Therefore, I want to include the set of omitted variables representing the economic or climate conditions.

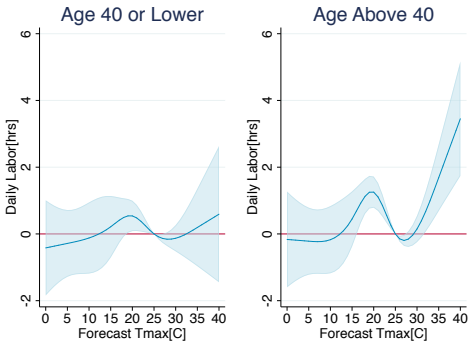
Figure 1.5: Heterogeneity Analysis with Two-Group Subsamples - Comparison of RMSE Marginal Effects



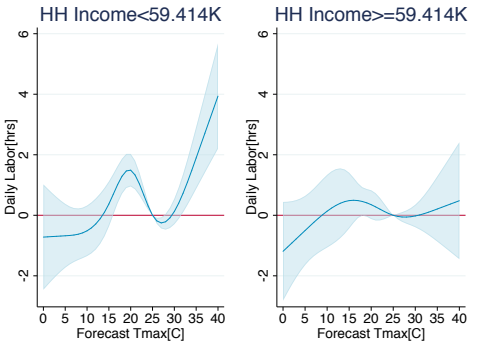
(a) Gender



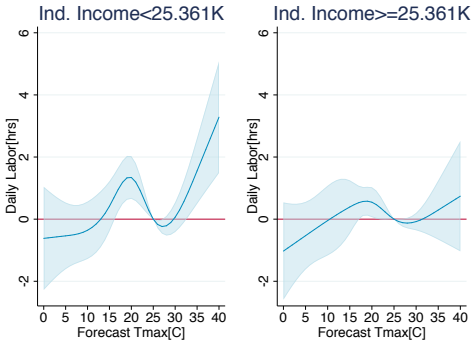
(b) Higher Degrees



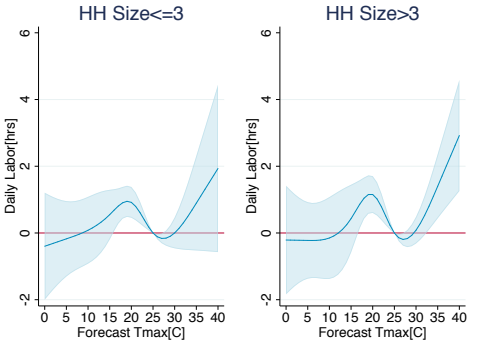
(c) Age



(d) Household Income

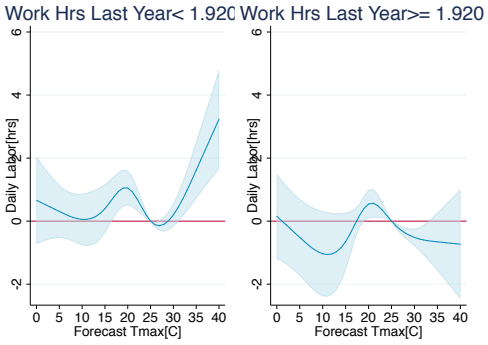


(e) Individual Income

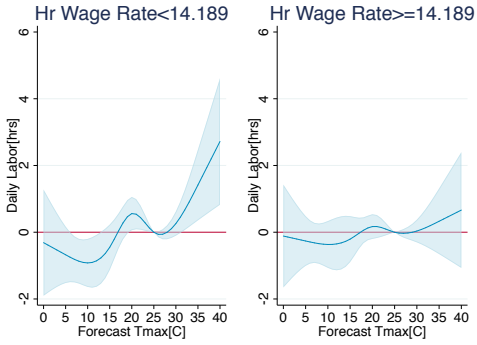


(f) Household Size

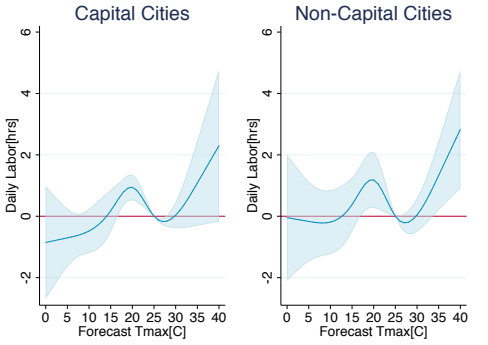
Figure 1.5, continued



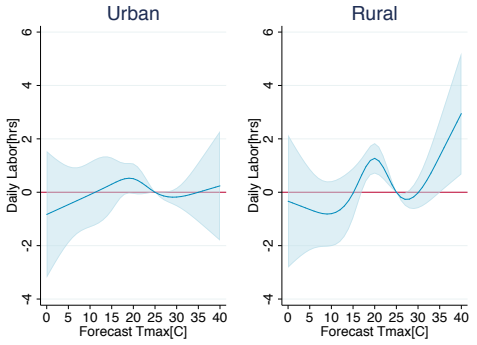
(g) Working Hours Last Year



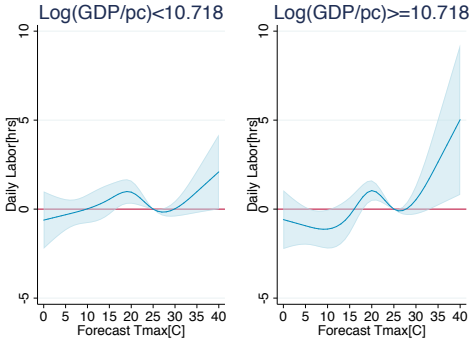
(h) Hourly Wage Rate



(i) Capital VS Non-Capital Cities



(j) Urban VS Rural



(k) Log GDP/pc

Note: Heterogeneity analysis is run with Eq. 1.3 where the two subgroups identified by 13 demographic and economic variables are allowed with different direct and $RMSE$ interacted marginal labor responses to forecasts. Marginal effects of $RMSE$ relative to $25C$ are estimated and compared by the variables defining these two-group subsamples. For continuous income and past labor variables (such as household income, working hours last year, GDP per capita), groups are divided by sample median.

First of all, I include the two most studied covariates representing economic and climate conditions, income (as higher income groups can be more flexible in adjusting their labor hours) and average climate (as residents living in hotter places can be more used to working under heat). Specifically, I take the income variable as the city-year GDP per capita in 2015 Yuan, and the climate variable as the city-year average realized daily $Tmax$[48]. Adding either or both covariates in Eq. 1.4, I estimate the marginal effects of $RMSE$ in Figure 1.6[49]. Overall, these marginal effects are little changed from the baseline in both magnitudes and statistical significance. Explanatory power increases with adjusted R-squared as expected, especially after adding the climate covariate of average $Tmax^{real}$. Arrays also maintain the baseline features (Figure 1.D.7). Therefore, the two major potential omitted variables do not alter my baseline estimates of labor responses to forecasts and forecast accuracy.

Next, I consider the sets of full economic and physical factors used in previous section (Table 1.C.1 and 1.C.2) possibly correlated with forecast $RMSE$. This is a saturated set, with both potential omitted variables like average income and climate, and those likely not determinants of labor response such as elevation[50]. Marginal effects of $RMSE$ are estimated for regression Eq. 1.4 including either or both sets of economic and physical factors as interactions[51] are shown in Figure 1.7. The rest two columns include city and season dummy indicators as the interaction covariates, meaning to allow the labor-forecasts response to be

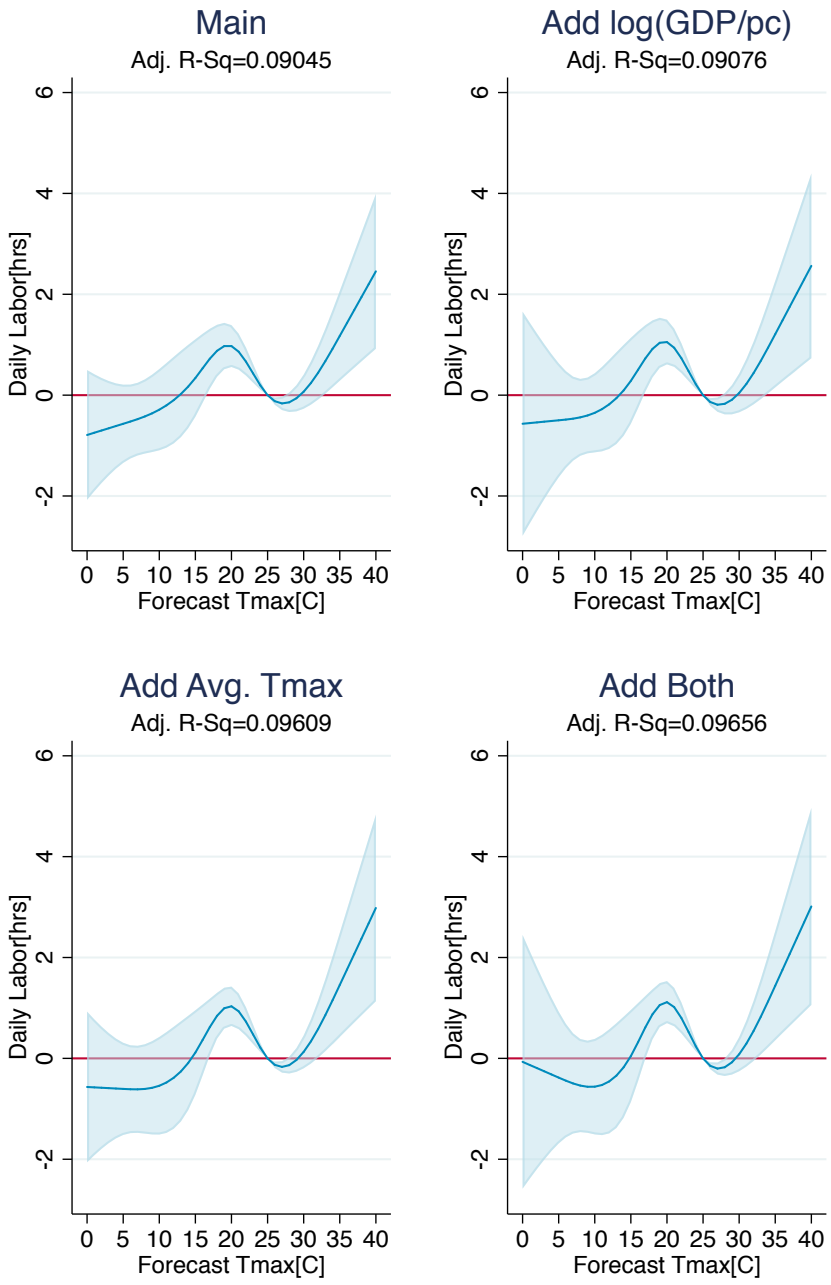
[48]There is also a version with previous five-year average as the historical climate variable, but regression results are little changed.

[49]Here I only take the common sample where neither covariates are missing, decreasing the sample size and adjusted R-squared from the baseline.

[50]Elevation is likely correlated with $RMSE$ through the forecast generation process, but it is unlikely to directly impact the labor-forecasts relationship.

[51]On a common sample with all these variables non-missing.

Figure 1.6: Regression with RMSE Interaction, Adding Income and Climate Interactions - Marginal Effects of RMSE



Note: Plots show the marginal effects of $RMSE$ interaction covariate relative to $25C$, estimated in the regression with double or triple interactions 1.4. Panels are the original main regression design (only RMSE interaction), adding income interaction, adding climate interaction, adding both income and climate interactions. For comparison, common sample is taken where neither income nor climate covariates are missing ($N = 10,457$). Income covariate is taken as the log of city-year GDP per capita, and the climate covariate is the city-year average realized daily $Tmax$ of 2011 and 2015.

city-specific or season-specific. Overall after adding economic and physical factors as interaction covariates, the positive significant marginal effects at hot end persist and even increase in magnitudes. Meanwhile, both magnitudes and statistical significance of the medium-cold forecasts positive marginal effects decrease. With the most controlled city-specific labor responses to forecasts, the positive marginal effects at hot and medium-cold again persist, though their magnitudes increase but statistical significance falls. The season-specific interaction controls post almost no change from the baseline, ruling out the impacts of seasonal changes in labor supply.

To summarize, the baseline results of this paper are not significantly changed by potential omitted variables to the interaction term of $RMSE$. The list of potential omitted variables is not exhaustive, so alternative empirical designs may be needed to further argue robustness to OVB. I then test with an instrumental variable regression using elevation[52] as the instrument for $RMSE$ interaction. Regression results are shown in Figure 1.D.8, still showing the hot and medium-cold labor decreases under low $RMSE$, though all estimates are quite noisy with extreme magnitudes[53].

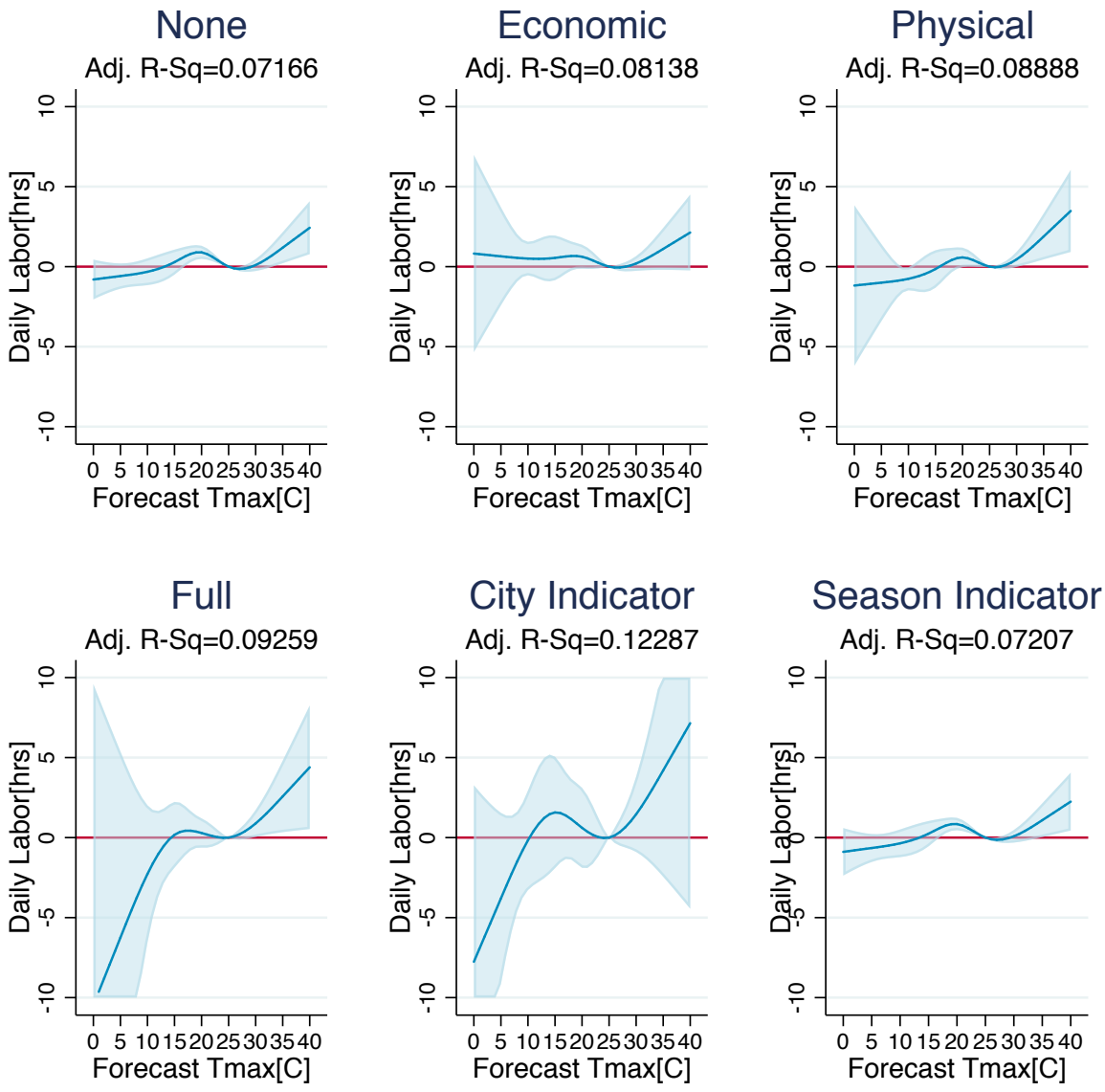
1.4.4 Other Robustness Checks

To further verify my baseline regression results of Eq. 1.2 in Figure 1.3, other robustness checks are addressed in this subsection.

[52]It correlates with climate conditions and possibly $RMSE$ but not other channels affecting the labor response to forecasts. In addition, I control for the yearly average real $Tmax$ for exogeneity assumption.

[53]Possibly due to the IV being weak by multiple tests (Cragg-Donald, Stock-Yogo, the non-significant correlations shown in Table 1.C.2).

Figure 1.7: Regression with RMSE Interaction, Adding Potential Omitted Variables Interactions - Marginal Effects of RMSE



Note: The plots show the estimated marginal effects of $RMSE$ interaction covariate relative to $25C$ from regression 1.4, adding full sets of economic or physical controls used in Section 3.1 as interactions. Panels are only RMSE interaction (main regression design), adding extra interaction covariates of all economic factors, all physical factors, both, adding city indicator interactions, and adding season indicator interactions. For comparison, common sample is taken across the panels where none of the economic or physical factors are missing (N = 9507).

Specification Tests for Functional Form $f(.)$ To test whether my baseline results are robust to the choice of functional form $f(.)$, I run the baseline with two alternatives described in Appendix 1.B, $5C$ bins and 3 knots linear splines. Results are compiled in Figure 1.D.9 together with the main restricted cubic spline specification, keep the half-year rolling $RMSE$ as interaction covariate. Overall, the labor responses estimated with the other functional forms are not statistically different for the main (linear spline is quite similar, bin is with much statistical uncertainties because of few observations in hottest bin). I also run the baseline for an alternative rolling window of four months (122 days), a minimum window size chosen with maximizing explanatory power (see Appendix 1.B). Results are presented in Figure 1.D.10, again baseline results remain robust, but statistical significance and estimate magnitudes drop for restricted cubic spline. These alternative functional forms and rolling window will also be included for sensitivity analysis in the next valuation step.

Alternative Fixed Effects My main specification uses city fixed effects though the smallest spatial unit is district, because I assume labors can work and live in different districts of the same city. On the other hand in order to increase estimation power and precision, I control month instead of the smallest temporal unit of week fixed effects. Replacing with more stringent fixed effect choices by districts (community) and by survey weeks (clustering alters with the same level as spatial fixed effects), Figure 1.D.11) shows that the baseline results do not change much.

Sample Restriction My labor sample includes extreme values of reported weekly labor hours ranging from 0 (no work at all) to 168 (24 hours all working per day). To check the

possibility of over or under reporting, I trim the top and bottom 1% weekly labor hours in my sample, dropping those working 0 hours and more than 84 hours (on average 12 hours per day) and leaving the trimmed sample size $N = 10,438$. Alternatively, I winsorize the weekly labor hours on the highest end by 84 hours and run with the original sample size $N = 11,012$. Regression results replicating Eq. 1.2 are presented in Figure 1.D.12, almost replicating the main results, especially the marginal effects.

Extra Linear Controls Considering the notable correlations of local economics and individual demographic characteristics with labor activities, I include additional linear controls of the city-year level log GDP per capita, log population and individual level age, gender and college degree indicators in $\mathbf{X}$. Figure 1.D.13 illustrates the robustness results, minimally changed from the main.

Realized Temperature Controls Though the main regression design exclude realized temperatures for multicollinearity concerns, I include the same 5 knots restricted cubic spline function for realized temperatures $Tmax^{real}$ as a linear control in Eq. 1.2 for robustness checks. Main results are again robust in Figure 1.D.14, when the large hot and medium-cold effects on labor decisions persist. Marginal effects are little changed from the main, but array estimates lose some statistical significance due to the high correlation between realized and forecast temperatures.

Limiting Temporal Variations of Perceived Forecast Accuracy To check whether including temporal variations of $RMSE$ affects the baseline results, I limit the day-to-day temporal variations of the $RMSE$ metric and replace with the yearly average city level fore-

cast $RMSE$ over three options, current year (2011 and 2015), previous year (2010 and 2014), and first observed forecast sample year (2010). In Figure 1.D.15, all three options replicate the arrays and marginal effects, while their regression R-squared decreases, suggesting more explanatory power contributed by the rolling $RMSE$ used for main analysis.

Different Forecast Accuracy Metrics I consider the following alternatives of the $RMSE$ metric as the indicator of historical weather forecast accuracy. Figure 1.D.16 replaces $RMSE$ with its two breakdown components (see 1.H.2 for details). Figure 1.D.17 runs with rationalized forecasts which have normalized forecast errors of mean zero[54]. Figure 1.D.18 assumes that individuals instead predict forecasts by themselves with an auto-regression model on historical realized temperatures[55]. Figure 1.D.19 has the regression run with splitted $RMSE$ separately interacted for cold and hot forecasts below and above $25C$[56]. Figure 1.D.20 applies maximum absolute error of forecasts over 183 and 30 days rolling windows[57]. Table 1.D.2 summarizes these alternative metrics with the main $RMSE$. From all these plots[58], my main results maintain throughout even though labor response estimates have decreased statistical significance and magnitudes. Except for the splitted $RMSE$, none of them has a higher adjusted R-squared than the main.

[54]Temperature forecasts are rationalized using the OLS estimates between realized and forecast temperatures on a rolling window of 183 days, for each city each date independently.

[55]The auto-regression model I use is AR(7), assuming individual predicts next day temperature forecasts with previous 7-day realized temperatures. Estimates are obtained from separate regression runs per city per day on a rolling window of 183 days.

[56]For this version, cold and hot $RMSE$ are summarized for forecasts over the previous 183 days below or above $25C$ only. Then they are separately interacted with cold or hot forecasts below or above $25C$ in the regression, so the result figure displays two marginal effects.

[57]Maximum absolute errors can be large especially for non-capital cities where forecasts are approximated at this stage, and these extreme values will be smoothed for average error metrics like the $RMSE$.

[58]Arrays are plotted towards different covariate minimums over the labor sample.

1.5 Valuation

1.5.1 General Model and Assumptions

The empirical section of this paper has addressed the impacts of increasing weather forecast accuracy (as represented by rolling $RMSE$ of city level daily $Tmax$ forecasts) on people's daily labor take-up decisions, in this case causing significant labor reductions under uncomfortable temperatures to work under (extreme hot, medium-cold). These results have implied that workers make use of forecast and forecast accuracy information in their labor decisions for day-ahead hours of work.

Based on this assumption, I propose a simple one-period utility maximization model under the general economic framework of labor-consumption trade-off with impacts from realized weather shocks. Consider the following setting. An individual decision maker has utility at time t dependent on one exogenous variable, the realized weather w, and two choice variables, labor l and consumption c. The decision maker needs to choose l and c in the night of previous day $t-1$, with the information of a forecast for the weather f (free-of-charge as a public good) and a forecast accuracy metric a. Utility is not realized until after the decisions are made in day t, when actual weather w is recognized.

Therefore, a typical worker completes the single-period utility maximization with uncertainty under his or her belief of the conditional distribution of real weathers $w|f, a$:

$$\max_{l,c} \mathbb{E}_w[u(c,l,w)|f,a]] = \max_{l,c}(\int_w u(c,l,w)p(w|f,a)dw)$$

$$s.t. -p_l l + c \leq I$$

Here consumption price is normalized to 1, therefore c stands for the total amount spent on any consumption goods. l is in unit of hours of working in a day, with hourly wage rate p_l. The budget of the day is I.

To simplify the solution, further simplifications are applied. Firstly, I assume no savings or fixed income investments, therefore the budget I is a fixed endowment to be consumed by the end of period. Under this assumption, the budget constraint binds and consumption can be substituted as $c = I + p_l l$. Secondly, for the utility function $u(c,l,w)$, I assume its functional form to be separable in l, c with a quadratic part of l and a linear part of c. Moreover, I assume the weather shock w affect worker's utility only through its interaction with the labor part, reflecting the disutilities of working under uncomfortable weathers.

With these additional assumptions, I define the utility function as a quadratic function in labor l and realized weather w[59]:

$$u(l,w) = \alpha l^2 + \beta(w)l$$

The parameter α is directly related to the labor return to wage rate. The function $\beta(w)$

[59]Note that since c is expressed linearly by l from the binding budget constraint, substitution merges the parameter p_l into the constant term of the interaction function $\beta(w)$.

expresses the sensitivity of worker's marginal utility of labor to realized weather w.

After simplifications, the worker's utility maximization problem can be solved with first order condition, assuming interior solution with budget I large enough[60]:

$$l^* = g(f, a) = -\frac{\mathbb{E}_w[\beta(w)|f, a]}{2\alpha}$$

Following this, the value of forecasts with accuracy a can be expressed as the objective function averaged across distribution of f:

$$V(a) = \mathbb{E}_f[\mathbb{E}_w[u(l^*, w)|f, a]] = \int_f \int_w u(l^*, w)p(f|a)p(w|f, a)dwdf$$

For analysis, I also define the intermediate value function as the average utility realized when the individual takes an action $l^* = g(f, a)$ and then counter the weather shock w:

$$\bar{V}(w, a) = \mathbb{E}_f[u(l^*, w)|w, a] = \int_f u(l^*, w)p(f|w, a)df$$

Using the Bayes' relationship and assuming the realized weather distribution $p_0(w)$ is independent of a, $V(a)$ is then the expected value of $\bar{V}(w, a)$ over the distribution of w:

$$V(a) = \int_w \bar{V}(w, a)p_0(w)dw$$

In later subsections, valuations of $\bar{V}(w, a)$ and then $V(a)$ will be approached.

[60] I has been dissolved as a constant term of the utility function, independent of the utility maximization problem.

1.5.2 Model Estimation with Baseline Regression Estimates

Since the individual decision maker can have non-rational belief of the conditional distribution $w|f,a$, their perceived objective functions could be different from the realized values. This paper will be presenting the realized instead of perceived objective value functions, using a novel estimation method without explicitly assuming on these beliefs.

To estimate the value functions $V(a)$ and $\bar{V}(w,a)$, I need to identify $\beta(w)$ for the utility function u (the parameter α can be referenced from labor elasticity, see the following discussion). Fitting the model with my data, I have the estimated equilibrium labor supply l^* as a function of the daily maximum temperature forecast f, and the perceived forecast accuracy metric $RMSE$ as a, corresponding to the analytical solution in my model $l^* = g(f,a) = -\frac{1}{2\alpha}(\mathbb{E}_w[\beta(w)|f,a])$. Assuming the decision maker would perceive that forecasts are "perfect" for a threshold of a^* such that when $a \leq a^*$, uncertainty dissolves and the decision maker believes that $f = w$ with 100% probability. Then the expectation term drops out under $a \leq a^*$:

$$g(w,a^*) = -\frac{1}{2\alpha}\beta(w) \Rightarrow \beta(w) = -2\alpha g(w,a^*)$$

Then the non-linear function $\beta(w)$ can be directly identified through the estimated labor response function $g(f,a)$ given a^*.

There are different choices of this threshold forecast accuracy a^*. In main valuation analysis, I will take $a^* = 1$, the minimum of the forecast sample covering all cities and all

days in 2011 and 2015. This value is selected assuming that people realize forecasts cannot get to their perfect points at $a = 0$, and they will be content with the current best observed forecasts. This value is also not far from the labor sample minimum $a = 1.4$ (to 1.d.p.), therefore making little extrapolation outside of the regression sample. In later subsection of sensitivity analysis, results with $a^* = 0$ (rational so only perfect forecasts are accepted) and $a^* = 1.4$ (no out-of-sample extrapolation) will also be presented.

Then I can propose the following steps for model estimation:

Step 1: Referencing α In my model, the scaler parameter α is the constant return of labor. I approximate this parameter $\alpha \approx -\frac{1}{2}(\frac{\Delta l}{\Delta p_l})^{-1} \approx -\frac{1}{2}(\eta \frac{l}{p_l})^{-1}$ with elasticity of labor supply, referenced as $\eta = 0.353$ (Li, 2016)[61]. In addition, $l = 7.496, p_l = 20.548$ are summarized from the labor sample used in this paper. In the end, I reference $\alpha = -3.882$.

Step 2: Identifying $l^* = g(f, a)$ **and** $\beta(w)$ The optimal labor decision $l^* = g(f, a)$ is directly estimated from the baseline regression Eq. 1.2. Relative to $25C$, this labor response function is deduced with reference labor $\bar{l} = 7.410$ hours summarized from the sample week with average daily forecast closest to $25C$ (see Appendix 1.E for details). As argued above, $\beta(w) = -2\alpha g(w, a^*)$ is estimated by this labor response function under the selection $a^* = 1$.

Step 3: Simulating $\bar{V}(w, a)$ $\bar{V}(w, a)$ is evaluated by averaging the estimated utility $u(l^*, w)$ for any pair of (w, a) across the normal simulations of $f|w, a \sim N(w - \mu, \sigma^2)$. I

[61]This paper estimates the elasticity quantity with choice-based conjoint analysis on the China Urban Labor Survey (CULS) with six cities (worker age 16-64, almost the same as my sample). I take its estimate for all workers in 2010, but other estimates for subgroups of labors will be used in sensitivity analysis.

simulate $\bar{V}(w,a)$ for the series of realized temperatures $w = 0,1,...,40$ under the $RMSE$ metric $a \in [1,4]$, fixing $\mu = 0$ (which is the sample minimum, see Appendix 1.H.2) and varying $\sigma = 1, 1.25, ..., 4$ (range from 1 to about the maximum over the labor sample).

Step 4: Aggregating to $V(a)$ I estimate $V(a)$ by taking the average of simulated $\bar{V}(w,a)$ weighted by a selected realized temperature distribution $p_0(w)$, which is the empirical distribution of $Tmax^{real}$ over the pooled sample of all 342 cities and all 730 days in 2011 and 2015. $V(a)$ is computed for $a \in [1,4]$ same as $\bar{V}(w,a)$. For display purpose, I normalize $V(a)$ subtracting the reference perfect forecast case $a^* = 1$ and inflate it by 365 days to annual level.

This model estimation method has the advantage of being simple, with minimum assumptions on workers' beliefs of realized weather distribution given forecasts and forecast accuracy. It is also directly linked to my baseline regression results from previous section, applying the estimated labor responses as Figure 1.2 Panel (c). Appendix 1.E gives the detailed descriptions of these valuation steps, including the generation of 95% confidence interval with Monte-Carlo draws from the baseline regression estimates.

1.5.3 Valuation Results

Main valuation results are summarized in Figure 1.8. Panel (a) shows $\bar{V}(w,a)$, featuring similar curves as the labor responses with lower values under hot or medium-cold temperatures. This is because the baseline regression has estimated significant decrease of labor under low $RMSE$, thus implying greater disutilities of working under these two temperature ranges. When a decreases, the $\bar{V}(w,a)$ curve flattens as expected, as value increases

with more accurate forecasts because under low $RMSE$ decision makers are able to choose closer-to-optimal labor responses to counter the weather shocks. The magnitudes of these improvement in $\bar{V}(w, a)$ with decreasing a is greater again over the two ranges, hot and medium-cold temperatures, where $RMSE$ has large magnitudes and statistically significant marginal effects in the baseline regression.

Next, Figure 1.8 Panel (b) illustrates $V(a)$ relative to $V(1)$, the value loss per person per year for workers facing non-ideal forecasts with $RMSE \in [1, 4]$. As expected, $V(a)$ is a decreasing function of a, with greater (less negative) $V(a)$ coming under greater forecast accuracy (smaller a). Starting from zero relative value under $a^* = 1$, to above 2800 2015 Yuan (446 USD)[62] for $a = 4$, the negative slope of the curve increases in magnitude with a, indicating that the marginal value of forecast accuracy is decreasing under lower $RMSE$. This is consistent with the diminishing return story, that when forecasts are already quite accurate, improving the accuracy is not as valuable as when they are with greater errors. The 95% confidence interval generated by Monte-Carlo (MC) draws is wide but never cross the zero line, affirming 5% statistical significance of these value estimates. The MC runs have greater magnitudes mean and median than the direct estimates, indicating that the distribution of $V(a)$ is skewed towards even greater values.

Linearly averaging from $RMSE = 4$ to $RMSE = 1$, the value gains per person per year per $1C$ decrease of $RMSE$ (the marginal value of forecast accuracy) is about 930 Yuan

[62] 2015 Exchange Rate 6.2837, source https://www.exchangerates.org.uk/USD-CNY-spot-exchange-rates-history-2015.html.

(148 USD) in 2015. Comparing with the sample average (Table 1.1), this marginal value is of considerable magnitude covering 26.6% of an average worker's monthly wage income. Referencing Bakkensen, Lemoine and Shrader (2022) who estimates the value of accurate temperature forecasts from the mortality response to instantaneous forecast errors in the US, their estimated marginal value of forecast accuracy is approximately 228 USD per person per year[63]. My valuation estimate is of same order of large magnitudes, but half of their value. This is sensible to the expectation that mortality is an aggregate measure with higher valuation results than partial valuations through behavioral responses like labor.

So far, my valuation has used assumed forecast accuracy. To compare with expenditures of the national weather forecast system in China, I conduct analysis to estimate the total national benefit from actual improvement of weather forecast accuracy. Allowing the $Tmax$ forecast yearly $RMSE$ for each city to change from their summarized values in 2011 to 2015 from my forecast dataset, I compare the changes in $V(a)$ for 276 cities with non-missing 2015 employed labor data under Figure 1.8 Panel (c)[64]. There are scattered spatial variations in these $V(a)$ value changes. Though most of the cities covered, especially those with higher populations, feature value gains due to their forecast $RMSE$ decreases, there are also inland cities with $RMSE$ increases leading to value loss.

[63]The Bakkensen, Lemoine and Shrader (2022) valuation number is quoted from the presentation at EPIC Workshop of the University of Chicago, November 1st, 2022. They conclude that the value per $1C$ decrease in forecast error standard deviation is 75.1 billion USD per year. To compare with my valuation results, I translate this number to the marginal value of $RMSE$ (in their data forecasts are unbiased with mean 0, same as my simulation) and divide 330 million US population.

[64]Simulations are run with parameters μ, σ empirically estimated by the annual forecast error mean and standard deviation for each city. I aggregate to $V(a)$ using the realized $Tmax$ empirical distribution over all days 2011 and 2015 for each city. Therefore, I allow the $p_0(w)$ distribution to vary by cities but not by years. Actually, the average realized $Tmax$ increases by about $0.7C$ over these 5 years.

Then summing with each city's 2015 employment number[65], Figure 1.8 Panel (d) summarizes a partial estimate of the annual national total value gain solely from changing the city level forecast accuracy in 2011 to that of 2015. During the 5 years when an average of 3.9% ($0.087C$) decrease of city-year forecast $RMSE$ happens over these cities[66], the society gains 25.3 billion 2015 Yuan (4.03 billion 2015 USD). This value is even greater by Monte-Carlo means at 46.1 billion 2015 Yuan (7.34 billion 2015 USD) since the distribution is positively skewed. The 95% confidence interval is $[18.0, 106.3]$ billion 2015 Yuan, making the estimate 5% statistically significant (actually all MC runs are positive from the histogram).

Overall, the magnitude of this national value gain estimate is quite large, consistent with the value of weather forecast estimates for other countries (mainly developed countries) discussed in introduction. The value of the entire national weather forecast system through contingent valuation is at least 46.5 billion Yuan in 2006 (Yuan, Sun and Wang, 2016)[67], and my valuation as a lower bound accounting only the impacts in labor sector and only the values from forecast accuracy improvement is over half of this total value. Comparing with the CMA's 2015 total reported expenditure of 26.3 billion Yuan, or the 22.4 billion Yuan allocated to meteorological services, my estimated annual benefit of 25.3 billion Yuan about covers their annual cost.

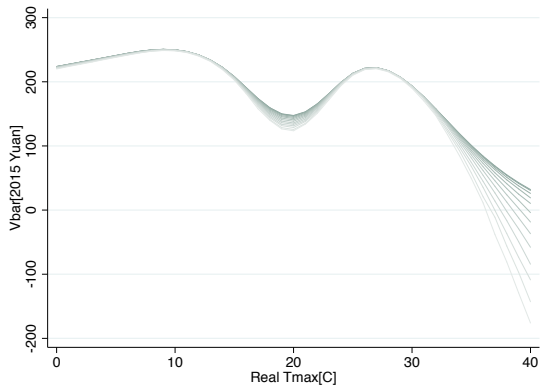
[65]Source China City Statistical Yearbook (CCSY) 2016, ignoring population growth. There are 364 million labor force covered in CCSY for 276 out of 342 cities, making this is a partial estimate. According to the national report of 774.51 million employed labors (Source: National Bureau of Statistics, http://www.stats.gov.cn/tjsj/zxfb/201602/t20160229_1323991.html, the city level employment data may be an underestimate by itself covering only 47% of the NBS's national record.

[66]This would feature a greater decrease compared with that whole sample with all 342 cities, summarized by Figure 1.2 Panel (d).

[67]Converted by World Bank GDP deflater, it is about 50.0 billion Yuan in 2015.

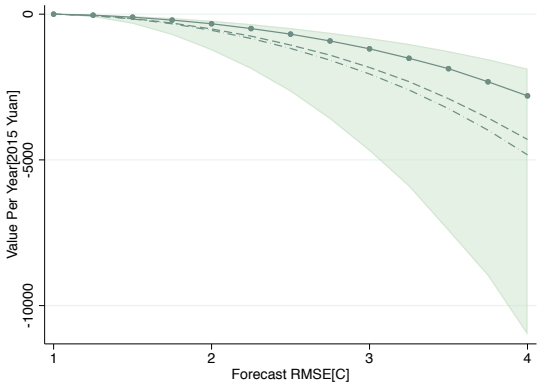
Figure 1.8: Valuation Results - $\bar{V}(w,a)$, $V(a)$ relative to $V(1)$, By-City Value Gain from 2011 to 2015 RMSE Changes, Total Value Gain from 2011 to 2015 RMSE Changes.

(a) Simulated $\bar{V}(w,a)$ Under Different a



Note: $\bar{V}(w,a)$ is the realized value of labors under weather shock w and forecast accuracy a. It is estimated by simulation assuming the realized distribution of $f|w,a \sim N(w-\mu,\sigma^2)$ for each $w = 0,1,...,40$. The mean forecasts error is fixed at $\mu = 0$, but forecast error standard deviation σ is allowed to vary. Each line indicates a case with $\sigma = 1, 1.25, ..., 4$, where lighter color represents greater σ. By definition, the $RMSE$ metric denoted as a in this case satisfies $a \approx \sqrt{\mu^2+\sigma^2}$. It ranges from 1 to 4 in this plot, so the darkest line represents the case $a = 1$ and the lightest line $a = 4$.

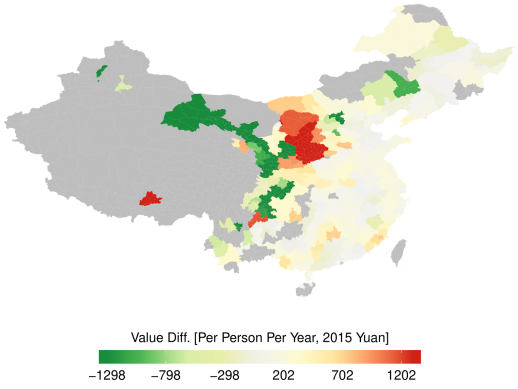
(b) $V(a)$ Relative to $V(1)$, Per Labor Per Year



Note: $V(a)$ is estimated by averaging the simulated $\bar{V}(w,a)$ values on the distribution of realized weather w. This distribution is taken as the empirical distribution of daily realized $Tmax$ over the all-city all-day sample of two years, 2011 and 2015. $V(a)$ is taken for various a under fixed $\mu = 0$, $\sigma = 1, 1.25, ..., 4$ so $a = \sqrt{\mu^2+\sigma^2} \in [1,4]$. The values are then subtract to relative of $V(1)$ (perfect forecast assumption under $a^* = 1$) and inflated to a year ($\times 365$). The plot shows result of the full valuation with direct estimates from the baseline regression in solid line/circle. Uncertainties are captured by 300 Monte-Carlo draws from the multinomial distribution of baseline regression estimates. The shade represents the 95% confidence interval. Dashed line is the MC median and dash-dot line is the MC average.

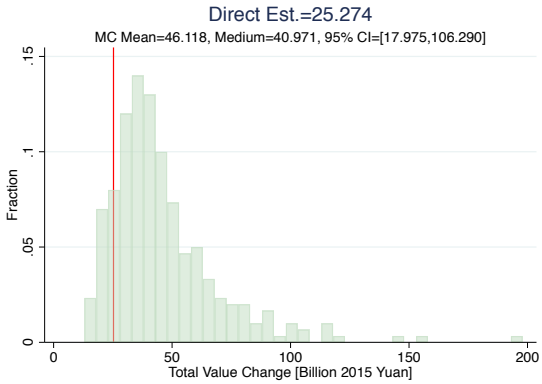
Figure 1.8, continued

(c) Value Gain in $V(a)$ by City from Their RMSE Changes 2011 to 2015



Note: Per person per year of the change in value $V(a)$ is computed for the scenario where each city has its yearly average $RMSE$ for $Tmax$ forecasts transformed from the value in 2011 to that in 2015. In the intermediate process, $\bar{V}(w,a)$ is simulated with the $Tmax$ forecast error mean and standard deviation (μ,σ) summarized over each year and each city. Note that $RMSE$ is determined by $a=\sqrt{\mu^2+\sigma^2}$. $V(a)$ is estimated per person per year averaging with the time-invariant realized temperature distribution for each city's daily $Tmax$ summarized over the all-day sample of two sample years, 2011 and 2015. For each city, the difference in $V(a)$ estimates for forecast accuracy 2011 and 2015 is evaluated and those with non-missing employment data in 2015 is illustrated in this map.

(d) Total National Value Gain from City Level RMSE Changes 2011 to 2015



Note: Total national value gain is obtained by summing the $V(a)$ changes by cities multiplying with city level employment in 2015 (source China City Statistical Yearbook), ignoring population growth. This is a partial estimate as total employment covers approximately 47% of the country's labor population in 2015 (source National Bureau of Statistics). The histogram features the distribution of the estimation results with 300 Monte-Carlo draws from the multinomial distribution of baseline interactive regression estimates. The vertical line indicates the total value gain with direct estimates of the regression.

1.5.4 Sensitivity Analysis

In this subsection, sensitivity analysis will be performed for the valuation of $V(a)$.

Different References of Scaling Factor α Estimates for the elasticity of labor supply differ by methodologies and different sample selections. To test the sensitivity of my results with different labor elasticity reference, I take the minimum and maximum elasticities by subgroups of education levels in Li (2016), at $\eta = 0.135$ (primary school degree or lower) and $\eta = 0.567$ (college degree or higher)[68]. These two elasticities convert to the parameter $\alpha = -10.152$ and $\alpha = -2.417$. Plotting for comparison in Figure 1.F.1, I can see that the main valuation result of $V(a)$, by method design, is just scaled up by the factor of 2.6 or down by the factor of 1.6.

Alternative Assumptions of a^* As argued previously, the definition of a^* can vary by assumptions of people's belief of "perfect" forecasts. To test the sensitivity with this assumption, I take alternative a^* at $a^* = 1.4$ (the labor sample minimum), and $a^* = 0$ (strictly rational with ideal forecasts of no errors). Valuation results of $V(a)$ relative to the corresponding $V(a^*)$ with confidence intervals are presented under Figure 1.F.2. By the average value gain per person per year per $1C$ decrease of $RMSE$[69], $a^* = 1.4$ decreases the valuation by only 1%, supporting that a little extrapolation outside the labor sample does not affect the main results. The $a^* = 0$ case increases this marginal value of $RMSE$ by up to 34%, showing the significant changes when extrapolation extends to the unrealistic case

[68]This is still a rather narrow range. Lower labor elasticities are expected among floating population (Chen and Zhu, 2021) and for later years (McClelland and Mok, 2012). In this case, floating population can have elasticity below 0.1 and the 2001 elasticity from the same paper can go above 0.97.

[69]The simulation range of a in each plot is different, where the minimum is taken at a^* and increases with steps of 0.25 until 4.2.

of $RMSE = 0$.

Change of Baseline Regression Specification To test the sensitivity with the estimated labor response function to forecasts, I compile the $V(a)$ estimates for three functional forms of the baseline regression (bins, restricted cubic spline, linear spline) with two $RMSE$ rolling windows (half a year and four months) addressed in previous empirical robustness section in Figure 1.F.3. The main valuation specification of restricted cubic spline with half-year rolling $RMSE$ actually gives conservative estimates compared with bin and linear spline, but the four-month window generates only half the original relative $V(a)$ in magnitudes. But overall, different specifications and rolling windows give same orders of magnitudes $V(a)$, large in real life and statistically significant at 5%.

Simulation Choice For simplification in computation, the main valuation uses a simulated range of a by fixing μ and varying σ. To verify the robustness of this simulation choice, I run the valuation instead on 0.25×0.25 grids with pairs of (μ, σ) extracted directly from the forecast data sample[70]. Results are summarized in Figure 1.F.4, where the simulation approximation represented by the lines mostly trace the scattering points estimated by the observed forecast accuracy grids. Therefore, my simulation process gives relatively good approximation to the real forecast accuracy distribution.

Other Realized Temperature Distributions In the estimation process, $V(a)$ as the average value of forecast accuracy is affected by the realized temperature distribution $p_0(w)$.

[70] I restrict the test to both μ and σ in the range of grids centering at $0, 0.25, ..., 4$, and run this for both labor sample and the all cities all days sample 2011 and 2015. Here a^* are taken as the minimum from these set of grid point representations of forecast data, approximately but not identical to 1.4 for labor sample and 1 for all-city sample.

For sensitivity checks, I allow two alternative distributions of $p_0(w)$, for the hot city of Haikou (the most south provincial capital city), and the cold city of Harbin (the most north provincial capital city)[71]. Results compiled in Figure 1.F.5 show that $V(a)$ does not change drastically (7.1% for Haikou and -21.4% for Harbin). From the breakdowns, valuation is significantly contributed by the medium-cold realized temperature cohort $[15C, 25C)$, because these temperatures come with greater frequency in all three distributions. The hotter city has greater magnitudes of $V(a)$ and the colder city smaller values because of the contribution from greater share of hot temperatures above $30C$. These observations illustrate the possible increasing value of accurate weather forecasts contributed by more frequent and hotter temperatures, which is motivating for future works related to climate change.

Implied Overreaction Belief to Inaccurate Forecasts In this paper, the labor responses estimated by the baseline regression and used for main valuation do not produce the highest realized utilities compared with two other labor responses not responding to forecast accuracy (Figure 1.G.1). This implies that people in my model are not making optimal labor decisions based on rational belief of forecasts. Actually, they may be putting much smaller weights on uncomfortable temperatures than they should be if they take the perceived $RMSE$ rationally[72]. These implied overreaction to not-so-inaccurate forecasts motivate a non-parametric model assuming individuals impose inflated conditional standard deviation of the normal belief $w|f, a$. This analysis detailed in Appendix 1.G concludes that it is unlikely decision makers form completely rational belief about forecasts with forecast

[71] Empirical distributions are taken for these cities of all days in 2011 and 2015. In addition, I breakdown the valuation of $V(a)$ by $5C$ bins of $Tmax^{real}$ to observe the contribution of different temperature bins.

[72] For example, the baseline regression shows an average forecasts error of $RMSE = 3C$ would completely eliminate any negative labor response even under extreme heat $Tmax^{forecast} = 40C$, while if rationally accessed people should respond like $Tmax^{real} \geq 30C$ and significant labor reduction shall still be observed.

accuracy. Overreacting as if there are greater uncertainties in forecasts would match the main empirical estimates, but it does not have sufficient evidence either in statistics or in reality.

1.6 Conclusion

This paper tries to estimate the economic value of weather forecasts, a common and popular scientific forecasting provided as public good in the modern world, with novel empirical microeconomic approach of revealed preference method. My research makes use of a unique video transcription panel dataset of daily weather forecasts in China, the largest developing economy with a great population of workers sensitive to climate shocks adapting to forecast information. My main empirical results suggest that labor decisions of hours worked per day respond to day-ahead temperature forecasts, and the medium-run average historical forecast accuracy determines whether these responses are significant. To be more specific, higher accuracy temperature forecasts induce decrease in labor under uncomfortable hot and medium-cold temperatures, but lower accuracy forecasts do not lead to significant changes in labor responses. With a single-period utility maximization model that incorporate the empirically estimated labor responses to forecasts and forecast accuracy, I evaluate a large marginal value of weather forecast accuracy in China such that the labor sector alone generates about enough annual partial social welfare to cover the annual governmental investments in its national weather forecasting system, from improving city level forecast accuracy by an average 3.9% from 2011 to 2015.

This project has verified the significance of maintaining a modern weather forecasting system that can provide accurate weather forecasts as a public good. Meanwhile, there remains ambiguities need to be addressed in future researches. Firstly, the large magnitudes of labor responses and valuations estimated in this paper are uncommon in previous literatures on labor-climate relationships. Alternative labor datasets should be used for verification, such as real-time labor records from specific industries (like Uber drivers), or repeated labor surveys with greater time frequencies. Secondly, direct proofs of public awareness to weather forecasts and historical forecast accuracy are required. This paper assumes people's belief of weather forecasts and their judgments with forecast accuracy based on the explanatory power of regressions, but people's belief system can be much more dynamic and complicated especially when information channels diversify. Possible directions in the future include to apply annual national survey data of weather forecasts from the government agency or use machine learning text analysis tools on social media posts to identify people's reactions towards weather forecast information. Thirdly, my study does not consider longer-run average impacts of forecasts or distinguish the influence from shorter-run forecast accuracy. For example, instantaneous forecast errors may affect people's final labor as they adjust their choices after realization, and people may be also planning inter-temporal substitutions so they work more under perceived accurate forecasts for comfortable temperatures. In that case, different regression models with lags of forecasts and instantaneous errors and multiple-period welfare analysis models should be considered.

The methodology of this paper can be applied to similar settings of identifying the eco-

nomic value of technology public goods through revealed preference. For the value of weather forecasts, my other works in progress have also found the positive value of accurate cold temperature forecasts on reducing road congestions and improving social media sentiments. To compile the aggregate social value of weather forecasts and to determine which sectors of the economy value accurate weather forecasts higher, other outcome variables common in literatures will be explored in the future, including industrial production and productivity, mortality and health risks, energy supply and demand, land and asset markets (Burke, Hsiang and Miguel, 2015; Severen, Costello and Deschenes, 2018). Last but not least, with the consideration of climate change in mind, the valuation model in this paper can be applied with future climate projections in order to estimate the extra value loss under accurate or inaccurate weather forecasts when temperatures get more extreme over this century. This would be important for the policy implication of how much improvements in forecasting technology should be helpful in the future world.

1.7 References

Barwick, Panle Jia, Shanjun Li, Liguo Lin, and Eric Zou. *From fog to smog: The value of pollution information.* No. w26541. National Bureau of Economic Research, 2019.

Burke, Marshall, Solomon M. Hsiang, and Edward Miguel. "Global non-linear effect of temperature on economic production." *Nature* 527, no. 7577 (2015): 235-239.

Carleton, Tamma A., Amir Jina, Michael T. Delgado, Michael Greenstone, Trevor Houser, Solomon M. Hsiang, Andrew Hultgren et al. *Valuing the global mortality consequences of climate change accounting for adaptation costs and benefits.* No. w27599. National Bureau of Economic Research, 2020.

Chen, Jie, Yufeng Zhu. "Estimating the Elasticity of Labor Supply: Understanding the

Recent Evolution of Chinese labor Market" *The Journal of World Economy* 8 , no. 516 (2021): 28-54.

Dee, Dick P., S. M. Uppala, Adrian J. Simmons, Paul Berrisford, Paul Poli, Shinya Kobayashi, U. Andrae et al. "The ERA-Interim reanalysis: Configuration and performance of the data assimilation system." *Quarterly Journal of the royal meteorological society* 137, no. 656 (2011): 553-597.

Downey, Mitch, Nelson Lind, and Jeffrey G. Shrader. *Adjusting to Rain Before It Falls.* Working Paper, 2021.

Fox, Glenn, Jason Turner, and Terry Gillespie. "The value of precipitation forecast information in winter wheat production." *Agricultural and Forest Meteorology* 95, no. 2 (1999): 99-111.

Garg, Teevrat, Matthew Gibson, and Fanglin Sun. "Extreme temperatures and time use in China." *Journal of Economic Behavior & Organization* 180 (2020): 309-324.

Goodarzi, Shadi, H. Niles Perera, and Derek Bunn. "The impact of renewable energy forecast errors on imbalance volumes and electricity spot prices." *Energy Policy* 134 (2019): 110827.

Graff Zivin, Joshua, and Matthew Neidell. "Temperature and the allocation of time: Implications for climate change." *Journal of Labor Economics* 32, no. 1 (2014): 1-26.

Guido, Zack, Sara Lopus, Kurt Waldman, Corrie Hannah, Andrew Zimmer, Natasha Krell, Chris Knudson, Lyndon Estes, Kelly Caylor, and Tom Evans. "Perceived links between climate change and weather forecast accuracy: new barriers to tools for agricultural decision-making." *Climatic Change* 168, no. 1 (2021): 1-20.

He, Guojun, Maoyong Fan, and Maigeng Zhou. "The effect of air pollution on mortality in China: Evidence from the 2008 Beijing Olympic Games." *Journal of Environmental Economics and Management* 79 (2016): 18-39.

Heal, Geoffrey, and Jisung Park. *Goldilocks economies? Temperature stress and the direct impacts of climate change.* No. w21119. National Bureau of Economic Research, 2015.

Hsiang, Solomon. "Climate econometrics." *Annual Review of Resource Economics* 8 (2016): 43-75.

Ivkovic, Zoran, and Narasimhan Jegadeesh. "The timing and value of forecast and recommendation revisions." *Journal of Financial Economics* 73, no. 3 (2004): 433-463.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). "Anomalies: The endowment effect, loss aversion, and status quo bias." *The Journal of Economic Perspectives*, 5(1), 193-

206. doi:10.1257/jep.5.1.193.

Katz, Richard W., and Allan H. Murphy, eds. *Economic value of weather and climate forecasts. Cambridge University Press*, 2005.

Lazo, Jeffrey K., Rebecca E. Morss, and Julie L. Demuth. "300 billion served: Sources, perceptions, uses, and values of weather forecasts." *Bulletin of the American Meteorological Society* 90, no. 6 (2009): 785-798.

Lemoine, Derek. *Estimating the consequences of climate change from variation in weather*. No. w25008. National Bureau of Economic Research, 2018.

Li, Yanan. *Estimating the Elasticity of Labor Supply in Chinese Cities. Economic Perspectives* 11 (2016): 68-78.

Martinez, Andrew B. "Forecast Accuracy Matters for Hurricane Damage." *Econometrics* 8, no. 2 (2020): 18.

McClelland, Robert, and Shannon Mok. "A review of recent research on labor supply elasticities." (2012).

Murphy, Allan H. "What is a good forecast? An essay on the nature of goodness in weather forecasting." *Weather and forecasting* 8, no. 2 (1993): 281-293.

Nurmi, Vaino, Adriaan Perrels, Pertti Nurmi, Silas Michaelides, Spyros Athanasatos, and Matheos Papadakis. "Economic value of weather forecasts on transportation: Impacts of weather forecast quality developments to the economic effects of severe weather." *EWENT FP7 project* (2012).

"Quality of ERA-Interim and comparison with other datasets: precipitation." Copernicus. 2017. https://climate.copernicus.eu/quality-era-interim-and-comparison-other-datasets-precipitation.

Rode, Ashwin, Rachel E. Baker, Tamma Carleton, Anthony D'Agostino, Michael Delgado, Timothy Foreman, Diana R. Gergel et al. "Labor disutility in a warmer world: The impact of climate change on the global workforce." (2022).

Rosenzweig, Mark R., and Christopher R. Udry. *Assessing the Benefits of Long-Run Weather Forecasting for the Rural Poor: Farmer Investments and Worker Migration in a Dynamic Equilibrium Model.* No. w25894. National Bureau of Economic Research, 2019.

Severen, Christopher, Christopher Costello, and Olivier Deschenes. "A Forward-Looking Ricardian Approach: Do land markets capitalize climate change forecasts?." *Journal of Environmental Economics and Management* 89 (2018): 235-254.

Shuman, Frederick G. "History of numerical weather prediction at the National Meteorological Center." *Weather and forecasting* 4, no. 3 (1989): 286-296.

Shrader, Jeffrey. "Expectations and adaptation to environmental risks." *Available at SSRN 3212073* (2020).

Shrader, Jeffrey G., Laura Bakkensen, and Derek Lemoine. "Fatal Errors: The Mortality Value of Accurate Weather Forecasts." (2022).

Teague, Kevin Anthony, and Nicole Gallicchio. "The evolution of meteorology: a look into the past, present, and future of weather forecasting." (2017).

Twersky, A., & Kahneman, D. (1974). "Judgment under uncertainty: Heuristics and biases." *Science*, 185(4157), 1124-1131. doi:10.1126/science.185.4157.1124 PMID:17835457.

Yuan, Huiling, Min Sun, and Yuan Wang. "Assessment of the benefits of the Chinese Public Weather Service." *Meteorological Applications* 23, no. 1 (2016): 132-139.

Zhang, Bing, F. Y. Zhai, S. F. Du, and Barry M. Popkin. "The China Health and Nutrition Survey, 1989-2011." *Obesity reviews* 15 (2014): 2-7.

1.A Weather Forecast Data Collection and Approximation

1.A.1 Capital City Weather Forecast Data Collection

This paper collects weather forecasts data for capital cities with the following steps:

1. Search: I want to find the full videos of the evening *Weather Forecast* program published by the official website of CCTV.com. Those are the real-time recordings of the program broadcast on TV, made available on the Internet hours later. By the time I started this project in 2019, the station stored their historical program videos on the old site tv.cntv.cn in five folders (2010-2012, 2013, 2014, 2015, 2016-current). I called the station in request of older videos (non-official sources of other video sites only have a few days of the program by individual blogger's preference), and they said earlier dates

could not be accessed due to copyright. Therefore, I can only collect weather forecasts from the year of 2010.

My purpose is to select only the evening forecasts (the one with consistently high viewership after the national news) and report the URL of these videos for downloading use. Into each folder, the videos are listed as below:

In the old website video folders, there are other news programs as well. So I first apply Bash commands curl and grep to collect all *Weather Forecast* videos, morning, noon and evening (example of 2015):

```
#!/bin/bash

output_loc=/Users/yuqisong/Desktop/Harris/WeatherURL/2015

foo () {
    local run=$1
    nohup curl -L http://tv.cntv.cn/videoset/VSET100224447532/page/${i} | grep -i
e '《早间\|《午间\|《晚间' > ${output_loc}/names_${i}.txt &
    #printf "hello ${i}"
}

for i in {1..117}
do
    foo "$run" &
done
```

The output .txt files contains title, image and hyperlink URL of all videos displayed on the webpage. Then I use R codes to further filter only the hyperlink URL http://tv.cntv.cn/... for evening *Weather Forecast* programs. I compile them by rows into batches of about 15 – 20 lines such that I can copy and paste onto a video downloading website.

For future forecasts data collection on the new website CCTV.com, I can directly search videos with keywords "Evening Weather Forecast" plus the year, making this process

much easier without the filtering commands.

2. Download: I batch download 15 – 20 videos per trial using the video consolidating-downloading site FLVCD.com. These videos are flash, making them harder to download using the other terminal commands.

   The downloaded videos are in .mp4 format. A day's full program is usually about 4 to 4.5 minutes, first half featuring the broadcaster discussing significant weather changes, usually related to natural disasters (for example, forest fires, extreme heats, sudden cold shocks). The second half just over 2 minutes are the next 24-hour temperature and categorical weather forecasts for 34 cities (example of January 1st, 2015, forecasting 6 cities in northern China):

   In each video, the broadcaster goes through the cities starting from Beijing in a fixed order, announcing in the format "City A, cloudy/sunny/rainy, $Tmin$ to $Tmax$ degree C". These lines are what my project wants to read and collect. There are other information in these videos, like the disaster forecasts first half of the program and the day after next forecasts only displayed on screen but not pronounced by broadcaster. These are useful information that may be applied to future researches.

3. Transform: Since I only need the next day forecasts announced by the broadcaster, I transform the .mp4 videos to .flac audio files by the ffmpeg command. To fit the wavelength requirement of transcription API, command sox is applied and these .flac files are converted to 1 (mono) channel, 16 bits. This process compresses the audio files to less than half of its original size. Code example:

```
#convert mp4 to flac by ffmpeg
#time estimate: 20min
for i in 201812*.mp4; do ffmpeg -i "$i" "../Audio/2018/${i%.mp4}.flac" >/dev/nul
l 2>&1 | grep -v configuration:; done

#convert flac to acceptable format
#time estimate: 1min
cd ../Audio/2018
for i in 201812*.flac; do sox "$i" --channels=1 --bits=16 "./Format/${i%.flac}.fl
ac" >/dev/null 2>&1 | grep -v configuration:; done
```

4. Transcribe: The .flac audio files with fitting wavelengths are then uploaded on Google bucket, and fed into the Google Cloud speech-to-text API. Introduction of the API is on https://cloud.google.com/speech-to-text, and I use the Linux command line package Google Cloud SDK 273.0.0. I specify the language and keywords with gcloud ml speech recognize-long-running and let this API transcribe my audio files to Chinese texts, output with gcloud ml speech operations describe to Java format (.json) transcripts. Code example:

```
#send to GOOGLE
# for later years, add the category of smog
for i in $foo; do gcloud ml speech recognize-long-running gs://ely_testing/$i --
language-code='zh' --async --sample-rate=48000 --hints=["北京","哈尔滨","长春","沈
阳","天津","呼和浩特","乌鲁木齐","银川","西宁","兰州","西安","拉萨","成都","重庆","贵阳
","昆明","太原","石家庄","济南","郑州","合肥","南京","上海","武汉","长沙","南昌","杭州"
,"福州","台北","南宁","海口","广州","香港","澳门","晴","阴","多云","大雪","中雪","小雪
","大雨","中雨","小雨","阵雨","阵雪","雨夹雪","霾","零下","摄氏度"] > "code_${i%.flac
}.json"; done
# run simultaneously, estimated run time: 2min

#obtain transcript output
for i in code_201812*.json; do var=($(jq -r '.name' $i)); gcloud ml speech opera
tions describe $var > z${i%.json}.json; done
```

The Google speech-to-text API has been a high quality transcription tool recommended by computer scientists. The transcription speed is fast, about 2 minutes only for up

to 30 audios. However, the transcription error rate is higher for non-English speeches. For future research, native Chinese speech-to-text API (e.g. from Tencent) should be an alternative option.

5. Clean: Finally, the JSON scripts are cleaned and read to .txt by `grep` and `jq`. I clean the .txt scripts containing the Chinese text reading of the *Weather Forecast* program with STATA. There are mainly two types of errors in cleaning. One is the regular transcription errors (for example, the city Macao is occasionally transcribed to "We") probably due to the limitation of the Google database of Chinese texts, which can be corrected systematically by STATA codes. The other is due to the quality of flash videos. Though professional TV broadcasters are required to be clear in speeches, being fast through over 34 cities within 2 minutes can make common ambiguities in speeches. The Google API is still not advanced enough to decipher all ambiguous speeches (even in English), thus resulting in random errors of transcription (for example in Chinese, "Ten" can be easily confused with "Four"). Those errors require me to go back and hand check with the original program videos.

   Roughly measured by the amount of correction codes, overall transcription errors do not necessarily distribute randomly across time and space. There are one or two broadcasters whose speeches are particularly hard to transcribe correctly (possibly due to a lower sound frequency), and for most broadcasters there are a couple of cities whose punctuations make their lines reading faster than the other cities. Moreover, winter temperatures below zero usually get more errors when the API seems to find transcribing the Chinese phrase "below zero" ambiguous. In order to ensure data quality,

I have run multiple sanity checks to ensure that transcription errors are eliminated as thoroughly as possible. For the future, further spot checking can be conducted by another researcher.

1.A.2 Non-Capital City Forecasts Approximation

For non-capital cities forecasts besides the 34 capitals reported in the popular national program, the imminent ongoing data work is in finding and transcribing existing videos of local province-wise weather forecasts (aired on local TV channels only right before the national news with much lower viewership). Ideally, I should construct a machine-learning based relationship between capital and non-capital city forecasts and historical realized weathers. But for now due to time and resource constraints, as an alternative, I approximate temperature forecasts for any non-capital city i by adjusting the forecasts of their provincial capital p with their difference of the monthly (m) average realized temperatures (source ERA-Interim) in the previous year $y-1$:

$$T^{forecast}_{itmpy} = T^{forecast}_{tmpy} + \frac{1}{N_{my-1}} \sum_{s\epsilon\{\text{year y-1, month m}\}} (T^{real}_{ismpy-1} - T^{real}_{smpy-1})$$

Where N_{my-1} is number of days in year $y-1$, month m. Categorized forecasts are taken as the same as their provincial capitals, which would be very rough approximations.

This data approximation process necessarily assumes that people living in non-capital cities still watch the *Weather Forecast* program (which is true according to viewership), and

they would infer their cities' weather forecasts from historical weathers and the forecasts of their capital cities (which is hard to address and only with anecdotal evidence). Another possible reason for this approximation is that the weathers and forecasts of nearby capital and non-capital cities are quite correlated under stable climate conditions. Proving either hypothesis would be difficult, so I conduct some statistical tests to indirectly support these assumptions.

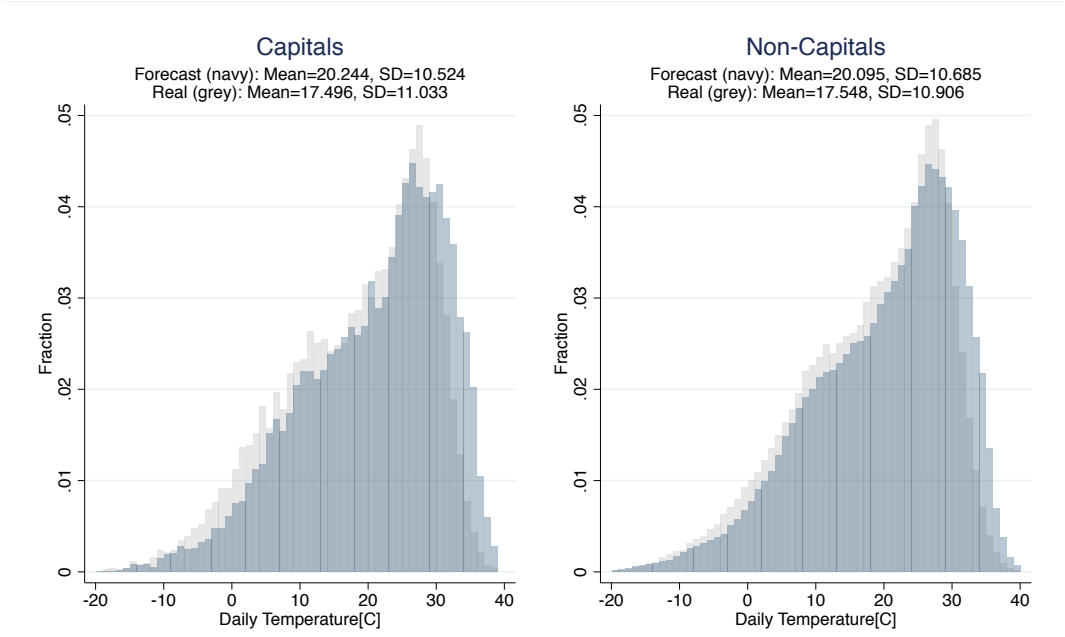
In particular, I propose the following hypothesis:

Hypothesis: Quality weather forecasts should be able to identify the similarity or discrepancy between different cities' temperature distributions.

For the argument that weather forecasts are "good enough", the main section has shown that forecast temperatures are usually higher than the real temperatures, but the second moments and shape of distributions have been quite similar. The same histograms are displayed in Figure 1.A.1 here for capital and non-capital cities under the addressed approximation process, for all days of the year 2011 and 2015. Again, forecasts are overestimating real temperatures but shapes of the distributions are more similar for either sample.

Then under this hypothesis, I assume that if weather forecasts in China is good enough and my approximations for non-capital forecasts reasonably track the actual forecasts, when one group of cities have different realized temperature distribution than the other group of cities, their forecast temperature distribution shall also be different. Vice versa, if the real-

Figure 1.A.1: Daily Maximum Temperature Distributions for Capitals VS Non-Capitals



Note: Navy=Forecast, Grey=Real; Sample covers all days of 2011 and 2015; Capitals contain 31 provincial capital or centrally-administered municipalities; Non-Capitals contain the rest 311 non-capital cities; In Labor Sample covers the city-day in the weekly labor survey only; Histograms are capped by a minimum of $-20C$ and maximum of $40C$.

ized temperature distributions between the two groups are the same, so shall their forecast temperature distributions. To test that, I perform the two-sample Kolmogorov-Smirnov test for both realized and forecast $Tmax$, between pooled sample of capital and non-capital cities under different temporal ranges in Table 1.A.1. For either of the three temporal ranges, the sample of capital cities have different real temperature distribution than the sample of non-capital cities with 1% KS test rejection. At the same time, the reported and approximated forecast distributions of capitals and non-capitals also reject all KS tests at 1%. As a matter of fact, if I separately test the samples on each year of 2011-2015, both real and forecast tests are still rejected at 10%. Therefore, the forecast approximation process I proposed for non-capital cities has been not unreasonable.

Table 1.A.1: Two-Sample Kolmogorov-Smirnov Test

Temporal Range	$Tmax^{real}$		$Tmax^{forecast}$	
	D	p-value	D	p-value
Full (2011-2015)	0.012	0.000	0.024	0.000
Two-Year (2011 and 2015)	0.013	0.002	0.027	0.000
Labor Sample	0.244	0.000	0.241	0.000

Note: Full time range covers all days of five years 2011-2015, the two-year sample only covers the survey years of 2011 and 2015; Labor sample further narrow down to days only covered by the labor reporting weeks; D is the KS test statistics, maximum distance between the capital and non-capital empirical CDFs, p-value is the KS test p-value; As both capitals and non-capitals have size $N > 50$, exact option is not required.

1.B Specification Choice

1.B.1 Regression Design

Firstly, I define the global regression (called "simple regression" in the main section) similar to previous literatures, where people's labor choice is estimated as a non-linear function

of forecast temperatures:

$$Labor_{ikt} = f(Tmax_{it}^{forecast}; \beta) + \gamma' \mathbf{X}_{it} + \epsilon_{ikt}$$

Here index k is per person surveyed, i is the city where the person resides, t is day the forecasts target. Controls $\mathbf{X}_{it}$ includes the real precipitation and its square, and month and city fixed effects (FE). Standard errors are clustered at city level. To match the frequency of the outcome variable, the weekly labor hours, all RHS variables are summed to weekly.

Then in this paper, the main interactive regression estimates the differential labor responses to forecast temperatures under different realized medium run forecast accuracy, expressed as the $RMSE$ of historical $Tmax$ forecasts on a rolling window size R:

$$Labor_{ikt} = f(Tmax_{it}^{forecast}; \beta_0) + f(Tmax_{it}^{forecast}; \beta_1) \times RMSE_{it} + \gamma' \mathbf{X}_{it} + \epsilon_{ikt}$$

And the control set $\mathbf{X}$ in addition to precipitations and fixed effects also include $RMSE$ as a linear control to complete the interactive regression design.

This specification choice section will first discuss the choice of non-linear function $f(.)$. Specifically, I perform specification choices for three non-parametric and semi-parametric functional forms, bins, restricted cubic spline, and linear spline. For bins I need to determine the temperature bins to include, and for the splines I need to select their knots. These selections should balance both the power of the regressions and precision of the estimates. Since interactive regressions depend also on the choice of rolling window R for the $RMSE$

metric, these functional form choice will weight higher on the performance of the global regression.

In the following subsections, I will discuss each of the three functional form individually, starting with bins. When evaluating the interactive regression, I take the rolling window size half a year, $R = 183$, and rolling window size four months (one third of a year), $R = 122$. The choice of window size R is then justified in Subsection B.6 after functional forms selections. The last subsection B.7 will run the non-parametric regression with binned $RMSE$ to robustness check the linear interactive design.

1.B.2 Bin Specification Choice

First of all, I run the relatively stringent non-parametric bin regression with $5C$ bin intervals $(-\infty, 0C), [0C, 5C), [5C, 10C)$, ..., $[35C, \infty)$. Basically, this means that function $f(.)$ includes series of dummies indicating whether the city-day temperature forecast is in each temperature bins. $1C$ bins are not considered due to precision issue, as there will be too many bins leading to noisy estimates. A reference bin $[20C, 25C)$ is omitted from the regression, taken as the comfortable temperatures for human (same as $Tmax^{forecast} = 25C$ in the main analysis).

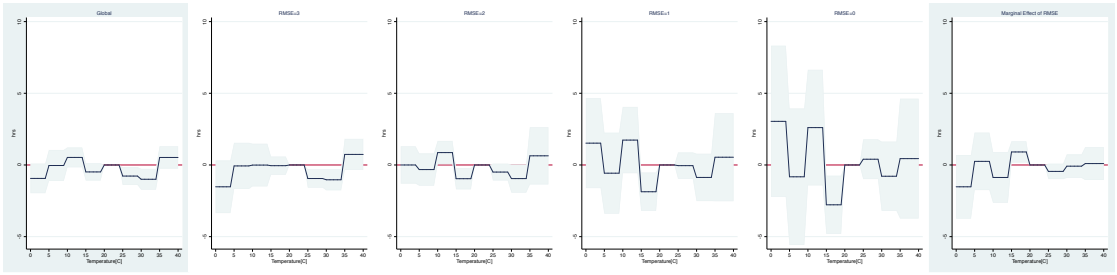
In previous literatures and in this paper, labor-temperature regressions usually focus on the hot side response above $Tmax = 25C$, because there is considerable increase in health risks working under heat. In that concern, I fix the hottest end bin at $[35C, \infty)$ and con-

sider merging on the coldest end bin. I test four coldest bin choices, $(-\infty, 0C)$, $(-\infty, 5C)$, $(-\infty, 10C)$, and $(-\infty, 15C)$. For each of these bin choices, I estimate both global and interactive regressions with $R = 183,122$, array plot summary in Figure 1.B.1 and 1.B.2. To compare the explanatory powers of the specifications, I summarize the adjusted R-squared and three trials of the average 10-fold cross validation out-of-sample (OOS) pseudo R-squared in Table 1.B.1.

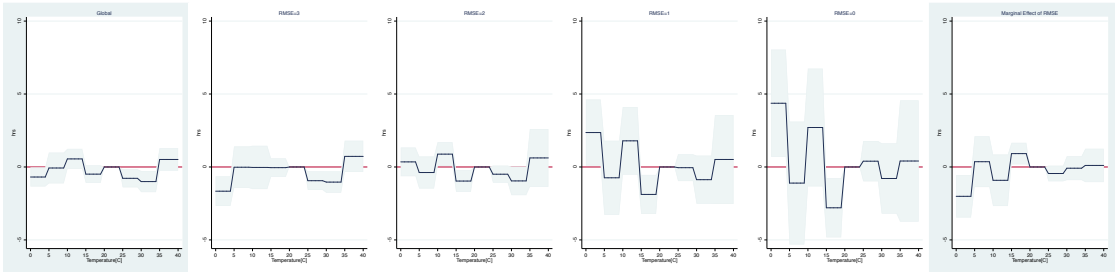
From Table 1.B.1, for global regressions, though adjusted R-squared roughly decreases with more merged bins down the rows, their OOS R-squared are close and fluctuating by different trials. The same applies to both interactive regressions, lending support to any bin combinations. Then looking into arrays in Figure 1.B.1 and 1.B.2, the hot and the mid-temperature responses are quite consistent throughout the choice of coldest bins, but cold end response can be quite fluctuating. For smoother cold end responses, I prefer the end bin choices $(-\infty, 10C)$ and $(-\infty, 15C)$. Going back to the R-squared table, the middle choice with cold bin $(-\infty, 10C)$ always have higher R-squared than $(-\infty, 15C)$, and even greater explanatory power than the other two choices $(-\infty, 0C)$ and $(-\infty, 5C)$ for some OOS trials.

Therefore, the final bin regression I choose is with 7 bins, $(-\infty, 10C), [10C, 15C), [15C, 20C)$, $[20C, 25C), [25C, 30C), [30C, 35C), [35C, \infty)$, and $[20C, 25C)$ as the reference omitted bin.

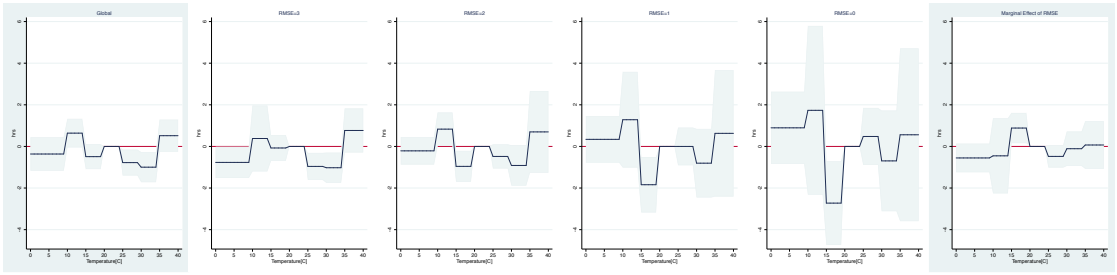
Figure 1.B.1: Bin Interactive Regression Rolling Window R=183



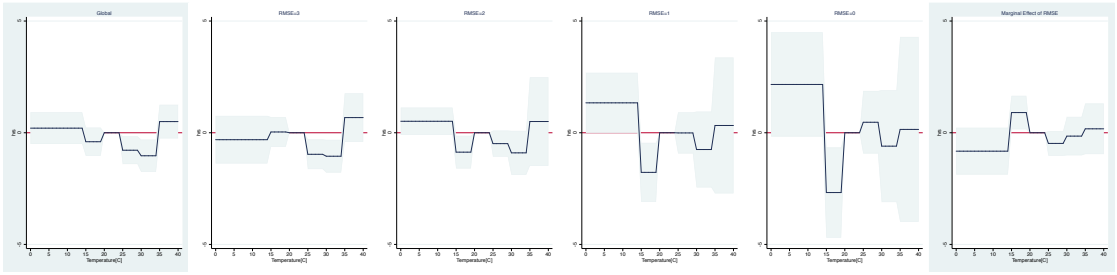
(a) Coldest Bin $(-\infty, 0C)$



(b) Coldest Bin $(-\infty, 5C)$



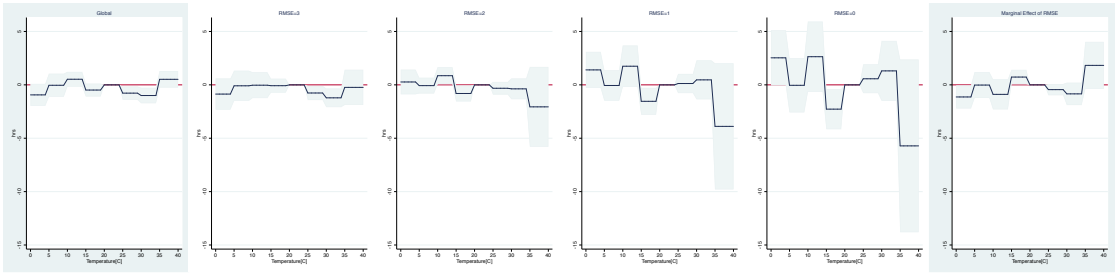
(c) Coldest Bin $(-\infty, 10C)$



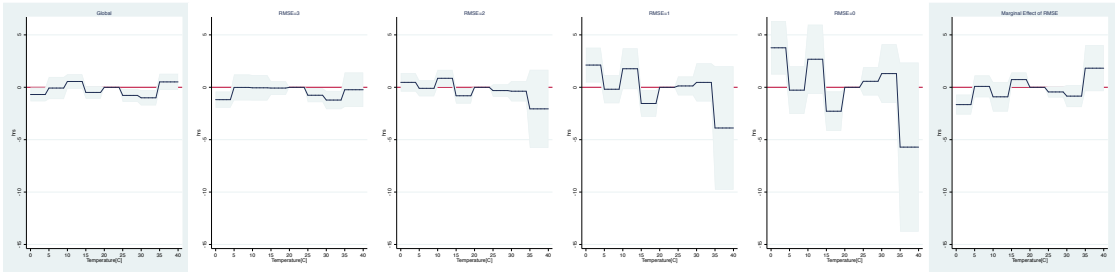
(d) Coldest Bin $(-\infty, 15C)$

Note: Left to Right, Global, RMSE decreases 3,2,1,0, Marginal effect of RMSE covariate.

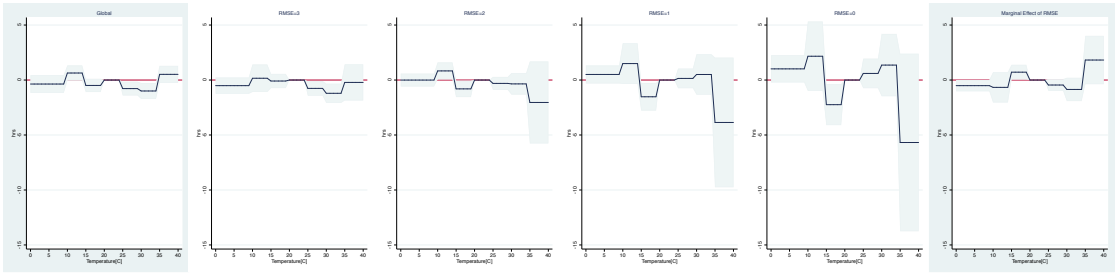
Figure 1.B.2: Bin Interactive Regression Rolling Window R=122



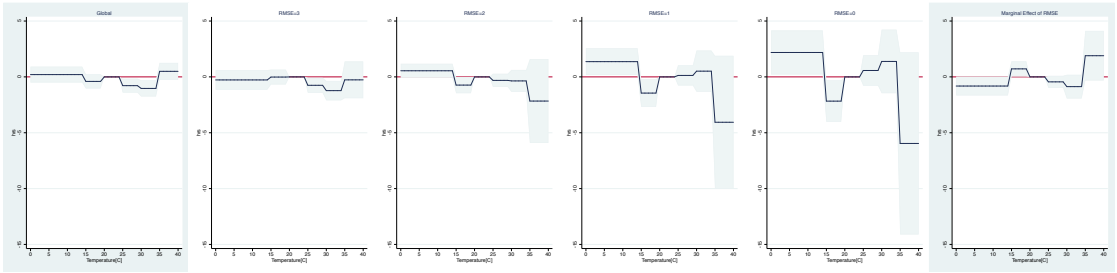
(a) Coldest Bin $(-\infty, 0C)$



(b) Coldest Bin $(-\infty, 5C)$



(c) Coldest Bin $(-\infty, 10C)$



(d) Coldest Bin $(-\infty, 15C)$

Note: Left to Right, Global, RMSE decreases 3,2,1,0, Marginal effect of RMSE covariate.

Table 1.B.1: Bin Specification Comparison

End Bin	Global R^2				Interactive R^2, R = 183				Interactive R^2, R = 122			
	Adj.	OOS (1)	OOS (2)	OOS (3)	Adj.	OOS (1)	OOS (2)	OOS (3)	Adj.	OOS (1)	OOS (2)	OOS (3)
$(-\infty, 0C)$	0.10193	0.09620	0.09747	0.10039	0.10791	0.10834	0.10028	0.10105	0.10230	0.10248	0.10147	0.10329
$(-\infty, 5C)$	0.10196	0.09665	0.09674	0.09554	0.10802	0.10846	0.10192	0.10267	0.10092	0.10279	0.10315	0.10255
$(-\infty, 10C)$	0.10182	0.09682	0.09699	0.09600	0.10777	0.10838	0.10233	0.10415	0.10124	0.10085	0.10196	0.10255
$(-\infty, 15C)$	0.10051	0.09508	0.09556	0.09469	0.10658	0.10779	0.09951	0.10092	0.10119	0.10262	0.10120	0.10100

Note: Adj. is the adjusted R-squared of regression; OOS R-squared is the average across 10-fold cross-validation out-of-sample R-squared.

Table 1.B.2: Restricted Cubic Spline Specification Comparison

Knots	Global R^2					Interactive R^2, R = 183				Interactive R^2, R = 122			
	Adj.	OOS (1)	OOS (2)	OOS (3)	RSS	Adj.	OOS (1)	OOS (2)	OOS (3)	Adj.	OOS (1)	OOS (2)	OOS (3)
$(5, 10, 15, 25, 30)$	0.09796	0.09376	0.09358	0.09260	0.77765	0.10284	0.10345	0.09730	0.09857	0.09699	0.09800	0.09642	0.09930
$(5, 10, 20, 25, 30)$	0.09782	0.09304	0.09361	0.09429	0.75090	0.10296	0.10347	0.09808	0.09882	0.09795	0.09775	0.09804	0.09924
$(5, 15, 20, 25, 30)$	0.09768	0.09388	0.09163	0.09120	0.73921	0.10270	0.10324	0.09673	0.09724	0.09821	0.09829	0.09783	0.09793
$(10, 15, 20, 25, 30)$	0.09754	0.09253	0.09153	0.09326	0.73586	0.10248	0.10305	0.09694	0.09770	0.09670	0.09702	0.09639	0.09882
$(5, 10, 15, 25, 35)$	0.09797	0.09121	0.09361	0.09288	0.73819	0.10363	0.10389	0.09902	0.09865	0.09801	0.09731	0.09795	0.09913
$(5, 10, 20, 25, 35)$	0.09799	0.09370	0.09337	0.09106	0.70338	0.10378	0.10405	0.09686	0.09774	0.09756	0.09804	0.09795	0.09749
$(5, 15, 20, 25, 35)$	0.09796	0.09192	0.09150	0.09332	0.69010	0.10354	0.10386	0.09735	0.09793	0.09910	0.09861	0.09793	0.09767
$(10, 15, 20, 25, 35)$	0.09788	0.09343	0.09198	0.09271	0.68679	0.10334	0.10369	0.09692	0.09613	0.09784	0.09652	0.09729	0.09823
$(5, 10, 25, 30, 35)$	0.09949	0.09423	0.09453	0.09287	0.57364	0.10464	0.10533	0.09939	0.09761	0.09948	0.09976	0.09978	0.09991
$(5, 15, 25, 30, 35)$	0.09959	0.09439	0.09412	0.09482	0.56463	0.10451	0.10510	0.09843	0.09930	0.09858	0.09931	0.09814	0.09976
$(5, 20, 25, 30, 35)$	0.09982	0.09342	0.09371	0.09623	0.54440	0.10455	0.10502	0.10040	0.10020	0.09903	0.09927	0.09873	0.09933
$(10, 15, 25, 30, 35)$	0.09954	0.09353	0.09475	0.09300	0.56423	0.10447	0.10503	0.09869	0.09916	0.09910	0.10031	0.09872	0.09925
$(10, 20, 25, 30, 35)$	0.09973	0.09496	0.09353	0.09560	0.54683	0.10452	0.10500	0.09850	0.09912	0.09909	0.09873	0.09825	0.09739
$(15, 20, 25, 30, 35)$	0.09966	0.09403	0.09507	0.09517	0.54805	0.10444	0.10495	0.09671	0.09841	0.09831	0.10072	0.09823	0.10019

Note: Adj. is the adjusted R-squared of regression; OOS R-squared is the average across 10-fold cross-validation out-of-sample R-squared; RSS means the root-mean RSS comparing the labor response relative to reference temperature curve with the selected bin regression.

1.B.3 Splines Specification Knots Choice

The bin regressions in previous subsection have confirmed that the functional form for $f(.)$ shall be non-linear. Next, I could explore the parametric non-linear functional forms for greater precision of estimates. In some less flexible parametric forms of $f(.)$ taking polynomials of orders $2, 3, 4$, it seems only when order of polynomials are higher than 3 I could efficiently replicate the non-linear labor response comparing with the bin specification choice of previous subsection. As a result, I would consider the minimum degree of freedom for parametric form to be 3. In this paper, I end up choosing restricted cubic spline with 5 knots and linear spline with 3 knots, both giving 4 degrees of freedom to the non-linear labor response function $f(.)$.

To determine the knots of splines, I propose the following selection rules:

1. Select from knots only at multiples of 5, i.e., knots are chosen from 7 values, $5C, 10C$, $15C, 20C, 25C, 30C, 35C$.
2. The reference temperature $25C$ has to be one of th knots.
3. There needs to be at least 1 knot at either side of reference temperature, i.e., at least one for $> 25C$ and one for $< 25C$.

That would give permutations of 14 options for restricted cubic spline with 5 knots and 8 options for linear spline with 3 knots.

Besides adjusted R-squared and OOS R-squared, I also estimate the error comparing global labor responses of splines with the non-parametric bin estimates. This method as-

sume the non-parametric bin estimates to be the "closest to real", and parametric splines should approximate as much as possible to the bin regression. Specifically, I estimate the residual sum of square (RSS) estimate in the following way for $Tmax^{forecast} = 0, 1, ..., 40$ ($Y = Labor, T = Tmax^{forecast}, \bar{T} = 25C$ is the reference temperature, b is bin specification, p is the parametric spline specification to be tested):

$$\Delta_b \hat{Y}(T, RMSE) = [f_b(T) + f_b(T) \times RMSE] - [f_b(\bar{T}) + f_b(\bar{T}) \times RMSE]$$

$$\Delta_p \hat{Y}(T, RMSE) = [f_p(T) + f_p(T) \times RMSE] - [f_p(\bar{T}) + f_p(\bar{T}) \times RMSE]$$

$$RSS = \sum_{T=0}^{40} [\Delta_b \hat{Y}(T, RMSE) - \Delta_p \hat{Y}(T, RMSE)]^2$$

For display purpose, I would summarize the root-mean of these RSS estimates as the average prediction error of parametric specification p comparing with non-parametric bin specification b. A smaller value of this will be preferred.

1.B.4 Restricted Cubic Spline Specification Choice

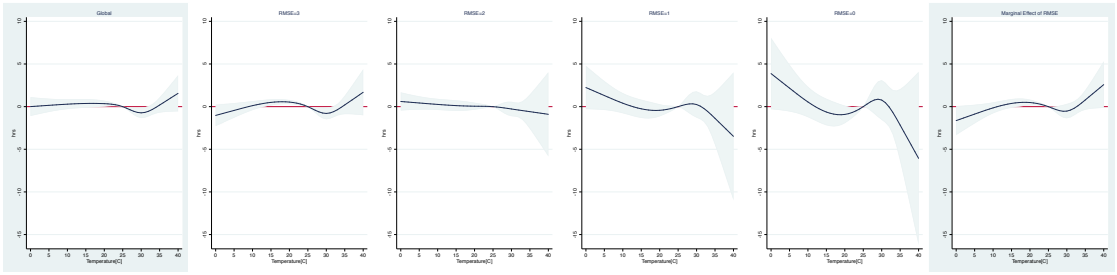
The R-squared, OOS R-squared and RSS estimates for 14 permutations of restricted cubic splines are summarized in Table 1.B.2. The 14 permutations are further separated into three groups by choices of hot knots being $30C$, $35C$ or both. From this table for global regressions, in general adjusted and OOS R-squared are higher and RSS lower for specifications with 2 instead of 1 hot knots. Among choices with 1 hot knot, the ones with $35C$ has greater R-squared and smaller RSS than the ones with $30C$. Similar observations apply for both

interactive regressions, where R-squared measures are overall in the decreasing order by hot knots $30C, 35C$, only $35C$ and only $30C$. Based on this table, I would consider the groups of permutations with hot knots $30C, 35C$ or only $35C$, excluding the hot knot with $30C$ only.

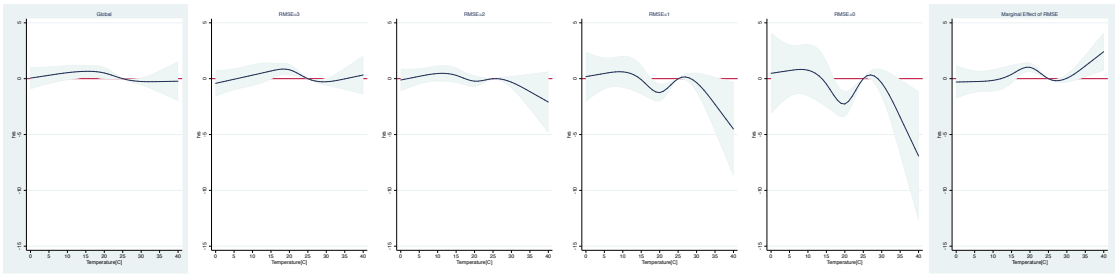
Then Figure 1.B.3 presents the array plots of interactive regression with rolling window $R = 183$, for two specifications with low RSS in their groups (first with both hotter knots at $30C, 35C$ and then with one hotter knot at $35C$). The response curves look similar, though the two-hotter knots specification (not only the one shows, but also all others in this group) estimates have wide confidence intervals towards the hottest end of the plots, both in arrays and in marginal effects. Therefore for precision, I instead prefer the one-hotter knot at $35C$ which would reduce over-identification and give statistical significant estimates at hot end. Within this group of 4 specifications, I end up choosing the combination $(5, 15, 20, 25, 35)$. It has slightly higher global RSS than its peers $(10, 15, 20, 25, 35)$ (the other two choices have greater RSS), but Figure 1.B.3 has shown them to be very similar. This choice with lower coldest knot has ensured higher R-squared especially for interactive regressions.

For the chosen specification with knots $(5, 15, 20, 25, 35)$, I also test the location of a mid knot $20C$ across $16C - 24C$. Though OOS R-squared fluctuates as before, both higher adjusted R-squared and lower RSS for global regressions have suggested the mid knot to be as close to $25C$ as possible. But if I look at the graphs in Figure 1.B.4, mid knot closer to the $15C$ knot would drive the cold end more wiggly, while closer to $25C$ would increase the extreme cold end response under low $RMSE$ to positive with a considerable magnitude. Otherwise important features like the hot end and medium-cold labor decrease at low

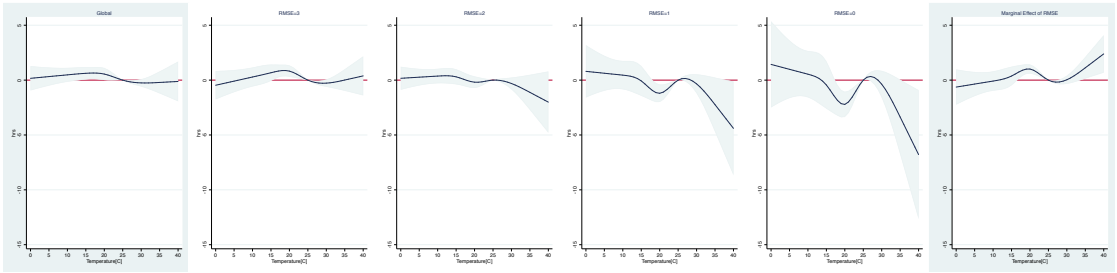
Figure 1.B.3: Cubic Splines Interactive Regression Rolling Window R=183



(a) Knots $(5, 20, 25, 30, 35)$



(b) Knots $(5, 15, 20, 25, 35)$

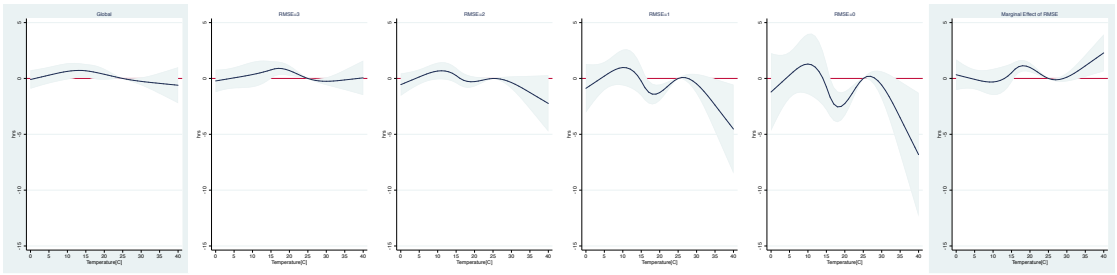


(c) Knots $(10, 15, 20, 25, 35)$

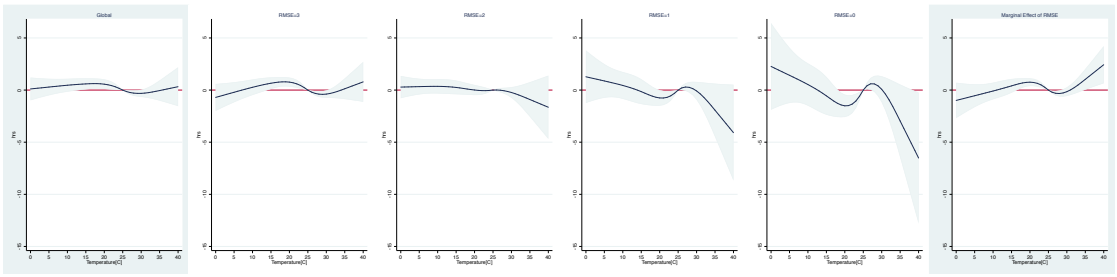
Note: Left to Right, Global, RMSE decreases 3,2,1,0, Marginal effect of RMSE covariate.

$RMSE$ all persist throughout various mid knot selections. As a result, I would keep the current selection of mid knot $20C$ at the middle of the range such that the cold end is neither wavy nor positive.

Figure 1.B.4: Cubic Splines Interactive Regression Rolling Window R=183



(a) Knots $(5, 15, 16, 25, 35)$



(b) Knots $(5, 15, 24, 25, 35)$

Note: Left to Right, Global, RMSE decreases 3,2,1,0, Marginal effect of RMSE covariate.

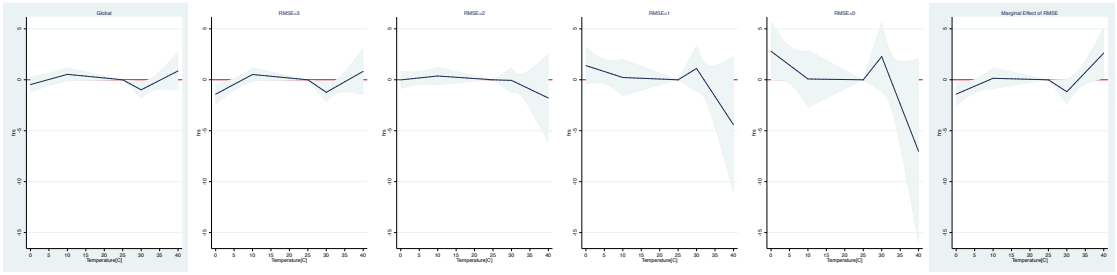
To summarize, the restricted cubic spline specification is chosen to be with 5 knots at $(5, 15, 20, 25, 35)$. This is also my main specification in the paper because it produces smooth response curves and statistically significant estimates.

1.B.5 Linear Spline Specification Choice

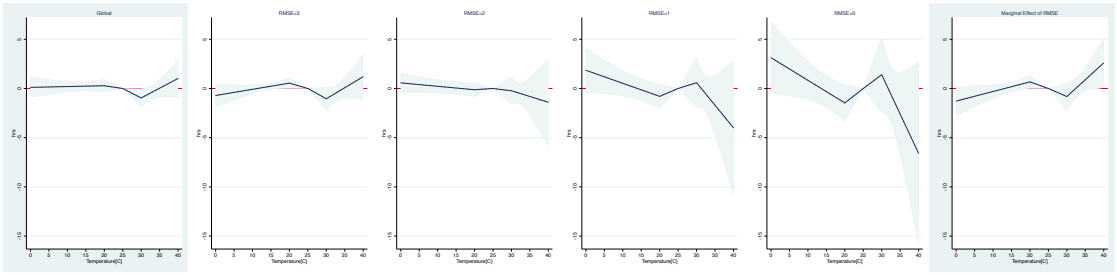
With similar agenda as the restricted cubic spline, I select linear spline specification among the 8 permutations listed in Table 1.B.3. Firstly for both higher adjusted R-squared and lower RSS of the global regressions, hot knot being $30C$ is preferred to $35C$. Among all permutations with hot knot $30C$, the position of the cold knot does not change much of the R-squared while OOS R-squared again fluctuates. The four specifications actually give similar negative but statistically insignificant responses at hot end, but as shown in Figure 1.B.5 the position of the cold knot affects the cold end and mid temperature response. The lower global RSS specification may be too wavy with positive and large in magnitude cold end responses for both array (under low $RMSE$) and marginal effect, but the medium-cold labor drop seen in the other specifications disappears. So to be consistent, I choose the combination with cold knot closest to $25C$, $(20, 25, 30)$, such the mid-temperature dip is preserved. This specification does not have lowest RSS or highest adjusted R-squared for the global regression, but its OOS R-squared is close or higher than peers especially for interactive regressions. Again to refine, I also test with altering the location of the cold knot by increment of $1C$, but again the regression curves and fitness metrics only vary little.

In the end, the linear spline specification is chosen to be 3 knots at $(20, 25, 30)$.

Figure 1.B.5: Linear Splines Interactive Regression



(a) Knots $(10, 25, 30)$



(b) Knots $(20, 25, 30)$

Note: Left to Right, Global, RMSE decreases $3, 2, 1, 0$, Marginal effect of $RMSE$ covariate.

Table 1.B.3: Linear Splines Specification Comparison

Knots	Global R^2					Interactive R^2, R = 183				Interactive R^2, R = 122			
	Adj.	OOS (1)	OOS (2)	OOS (3)	RSS	Adj.	OOS (1)	OOS (2)	OOS (3)	Adj.	OOS (1)	OOS (2)	OOS (3)
(5, 25, 30)	0.09982	0.09493	0.09491	0.09509	0.48629	0.10389	0.10487	0.09817	0.09804	0.09836	0.09992	0.09791	0.09876
(10, 25, 30)	0.10018	0.09664	0.09524	0.09342	0.44650	0.10437	0.10516	0.09930	0.10005	0.09741	0.09939	0.09757	0.09960
(15, 25, 30)	0.09947	0.09389	0.09244	0.09453	0.45920	0.10467	0.10532	0.09857	0.09993	0.09958	0.09803	0.09985	0.09960
(20, 25, 30)	0.09920	0.09356	0.09562	0.09339	0.48583	0.10436	0.10518	0.09880	0.10009	0.10010	0.09918	0.09833	0.09977
(5, 25, 35)	0.09866	0.09470	0.09347	0.09330	0.91852	0.10018	0.10260	0.09465	0.09704	0.09431	0.09679	0.09453	0.09739
(10, 25, 35)	0.09925	0.09369	0.09447	0.09474	0.87706	0.10132	0.10309	0.09494	0.09643	0.09310	0.09738	0.09569	0.09601
(15, 25, 35)	0.09851	0.09342	0.09356	0.09290	0.90026	0.10227	0.10369	0.09423	0.09763	0.09652	0.09859	0.09726	0.09885
(20, 25, 35)	0.09837	0.09308	0.09254	0.09296	0.95929	0.10240	0.10379	0.09720	0.09839	0.09640	0.09984	0.09728	0.09829

Note: Adj. is the adjusted R-squared of regression; OOS R-squared is the average across 10-fold cross-validation out-of-sample R-squared; RSS means the root-mean RSS comparing the labor response relative to reference temperature curve with the selected bin regression.

1.B.6 Rolling Window Size Choice

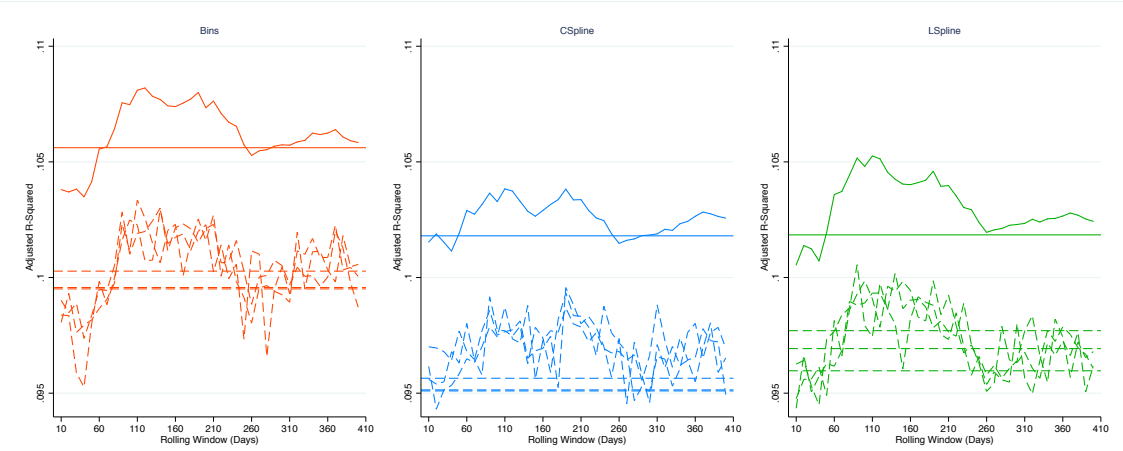
After selecting the non-linear labor response functional forms, I want to verify the choice of the size of rolling window for $RMSE$. For that purpose, I run the main interactive regression and record their adjusted R-squared together with three trials of the 10-fold cross-validation out-of-sample R-squared under different rolling window size R in Figure 1.B.6. Results are shown for all rolling windows from 10 to 400 days with an increment of 10. The horizontal lines feature the comparison taking the sample city-year forecasts $RMSE$ (2011 or 2015 per city) as the covariate instead. Overall, rolling window $RMSE$ delivers higher R-squared and OOS R-squared (though smaller with more fluctuations) mostly higher than the city-year $RMSE$ reference on windows $R = 60 - 200$ days, highlighting the increased explanatory power with possible adaptation to rolling estimation of the forecast accuracy metric $RMSE$. These R-square measures peak at about $R = 120 - 190$ days, translating to the medium run rolling $RMSE$ on a window of $4 - 6$ months, consistent for all three selected functional forms. As expected, the non-parametric bin specification outputs higher R-squared. Meanwhile, the two other semi-parametric spline specifications come with fairly similar levels of explanatory powers.

Further plotting the interactive regression marginal effects for different rolling windows in Figure 1.B.7, the estimates are not statistically different from one another across the columns. For bins, the very noisy hot end estimates have magnitudes decreasing to zero with larger rolling window size R. For restricted cubic spline, the reverse happens while rolling window closer to 100 gives smaller magnitudes and statistically insignificant positive

ME estimates at hot end. On the other hand, the marginal effect estimates for linear spline is the most stable across different rolling window sizes.

Based these explorations, I choose the optimal window size as $R = 183$, half a year, for my main restricted cubic spline specification. I also reserve the option at the lower end of the optimal window range, $R = 122$, or the window of four months, for robustness checks.

Figure 1.B.6: Explanatory R-Squared for Different Temporal Definition of RMSE

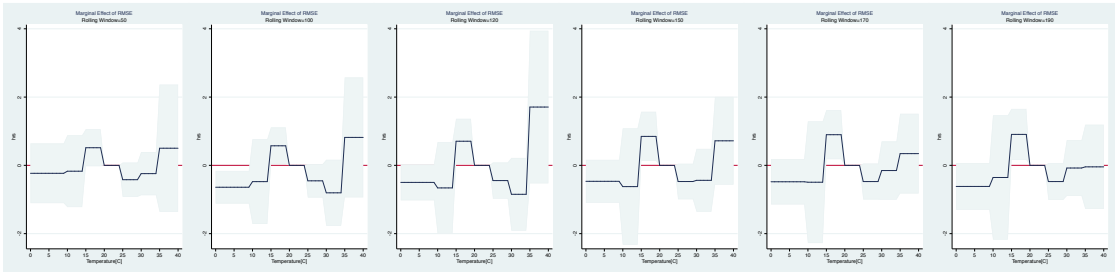


Note: Adjusted R-Squared (Solid) and 10-Fold CV OOS R-Squared (Dashed) are presented; Horizontal Line Indicates City-Year RMSE Choice.

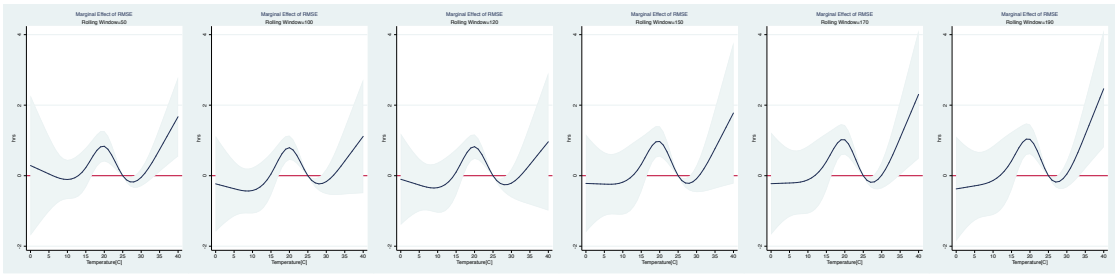
1.B.7 RMSE Linear Interaction Robustness Test

In the main section, labor response to forecasts and forecast accuracy is estimated with a parametric formula assuming linear interaction with the $RMSE$. To verify the robustness of this linearity design, I run the non-parametric interactive regression as follows:

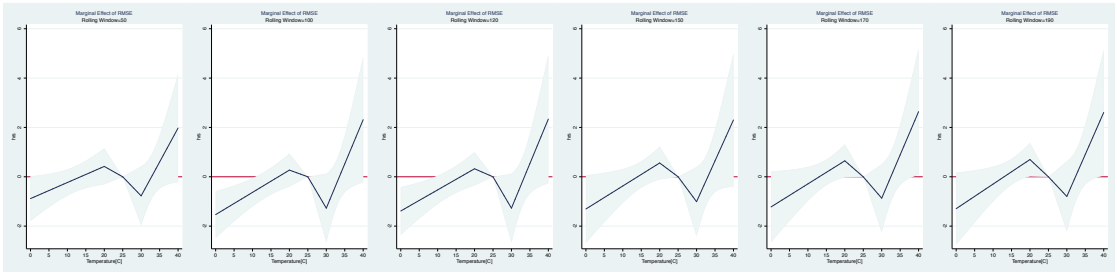
Figure 1.B.7: Interactive Regression Marginal Effect



(a) Bins



(b) Restricted Cubic Spline



(c) Linear Spline

Note: Left to Right, RMSE rolling window of 50, 100, 120, 150, 170 and 190 days.

$$Labor_{ikt} = \sum_{q} \mathbf{1}_{\{RMSE_{it} \in q\}} f(Tmax_{it}^{forecast}; \beta_q) + \gamma' \mathbf{X}_{it} + \epsilon_{ikt}$$

This is the regression estimating the non-linear labor response of labor to forecasts separately for different ranges of $RMSE$ denoted by q. Here the covariates set includes the series of dummies indicating the bins $RMSE$ is in, $\mathbf{1}_{\{RMSE_{it} \in q\}}$, instead of the linear covariate $RMSE$. Similarly, the independent linear control of $RMSE$ in the set $\mathbf{X}$ is replaced by these dummy variables as well.

I run this regression selecting q into three equidistance groups, $[1C, 2C), [2C, 3C), [3C, \infty)$, and results are presented under Figure 1.B.8. Overall, the results matches the main arrays estimated by the baseline regression design with $RMSE$ interacting linearly (Figure 1.3 Panel (c)) in both shape and magnitudes, preserving the decreasing labor responses for hot and medium-cold forecasts when $RMSE$ is in smaller bins. As expected, adjusted R-squared increases a bit for relaxing the linearity assumption, while estimates drop to statistical insignificance. Therefore, this non-parametric bin interactive regression justifies the robustness of my baseline regression design using $RMSE$ as a linear covariate for interaction.

1.C RMSE Variation Decomposition Regression Tables

Table 1.C.1: RMSE Variation Decomposition Regression with Economic Controls

	(1)	(2)	(3)
City Area [$10^4 km^2$]	0.243**	1.175***	1.326***
	(0.109)	(0.380)	(0.386)
GDP per Capita [10^8 2015 Yuan]	-0.057	0.006	0.003
	(0.086)	(0.059)	(0.056)
Population [10^9]	-0.004	-0.010*	-0.009*
	(0.006)	(0.005)	(0.005)
Road Area [$10^6 km^2$]	0.004	-0.005	-0.005
	(0.009)	(0.007)	(0.004)
Labor Force/Pop.	0.719	0.402	0.765*
	(0.517)	(0.417)	(0.412)
Unemployment Rate [%]	-0.038	-0.041	-0.041
	(0.054)	(0.061)	(0.058)
Share of Primary Industry [%]	0.025	0.016	0.016
	(0.019)	(0.021)	(0.020)
Share of Secondary Industry [%]	0.014	0.015	0.008
	(0.014)	(0.011)	(0.010)
Factors	Economic	Both	Both
Date FE	No	No	Yes
Observations	9507	9507	9507
Adjusted R^2	0.405	0.598	0.689
OOS R^2	0.405	0.597	0.688

Note: Dependent variable is the half-year rolling $RMSE$ for daily $Tmax$ forecasts; Standard errors in parentheses clustered by cities; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Factors indicate whether the regression include only economic or both economic and physical factors; FE indicates whether a date fixed effect is included in the regression; Sample include only the city-days overlapping in the labor sample with all economic and physical factors non-missing; OOS R^2 indicates the average across 10-fold cross-validation out-of-sample R^2.

Table 1.C.2: RMSE Variation Decomposition Regression with Physical Controls

	(1)	(2)	(3)
Water Resources ($10^8 m^3$)	0.004	-0.000	-0.002
	(0.003)	(0.004)	(0.004)
Greenland Area [$10^4 km^2$]	-3.503	15.023**	11.981**
	(2.188)	(5.623)	(5.523)
Elevation [km]	0.190	-0.504	-0.755
	(0.438)	(0.457)	(0.466)
No. Weather Stations	-0.196*	-0.171*	-0.125
	(0.098)	(0.091)	(0.082)
Area of Lakes [km^2]	-1.242	-0.649	-0.202
	(2.416)	(1.987)	(1.791)
Lenghth of Rivers [Deg]	0.006	-0.021*	-0.024**
	(0.008)	(0.011)	(0.011)
Annual Avg. $Tmax$ [C]	-0.090**	-0.019	-0.050
	(0.035)	(0.038)	(0.035)
Annual STD. $Tmax$ [C]	-0.190***	-0.173*	-0.254***
	(0.069)	(0.090)	(0.074)
Annual Avg. Precip. [mm]	-0.048	-0.055	0.045
	(0.068)	(0.062)	(0.064)
Factors	Physical	Both	Both
Date FE	No	No	Yes
Observations	9507	9507	9507
Adjusted R^2	0.489	0.598	0.689
OOS R^2	0.487	0.597	0.688

Note: Dependent variable is the half-year rolling $RMSE$ for daily $Tmax$ forecasts; Annual averages for $Tmax$ or Precip. are average real weathers of the current year; Standard errors in parentheses clustered by cities; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Factors indicate whether the regression include only physical or both economic and physical factors; FE indicates whether a date fixed effect is included in the regression; Sample include only the city-days overlapping in the labor sample with all economic and physical factors non-missing; OOS R^2 indicates the average across 10-fold cross-validation out-of-sample R^2.

Table 1.C.3: RMSE Variation Decomposition Regression with Economic Controls, Full Sample 2011 and 2015

	(1)	(2)	(3)
City Area [$10^4 km^2$]	0.165***	0.052	0.071
	(0.058)	(0.093)	(0.096)
GDP per Capita [10^8 2015 Yuan]	-1172.050***	-408.521*	-299.706
	(362.850)	(225.654)	(227.855)
Population [10^9]	-39.856*	-0.320	6.519
	(23.323)	(22.421)	(22.956)
Road Area [$10^6 km^2$]	-7800.821	-6632.905*	-6370.422*
	(5411.839)	(3646.315)	(3640.522)
Labor Force/Pop.	0.892*	0.625**	0.603**
	(0.457)	(0.299)	(0.292)
Unemployment Rate [%]	-0.013	0.001	0.001
	(0.030)	(0.027)	(0.027)
Share of Primary Industry [%]	-0.015	0.015	0.015
	(0.016)	(0.010)	(0.010)
Share of Secondary Industry [%]	0.003	0.016***	0.013**
	(0.009)	(0.006)	(0.006)
Factors	Economic	Both	Both
Date FE	No	No	Yes
Observations	156759	156759	156759
Adjusted R^2	0.188	0.493	0.521
OOS R^2	0.188	0.493	0.519

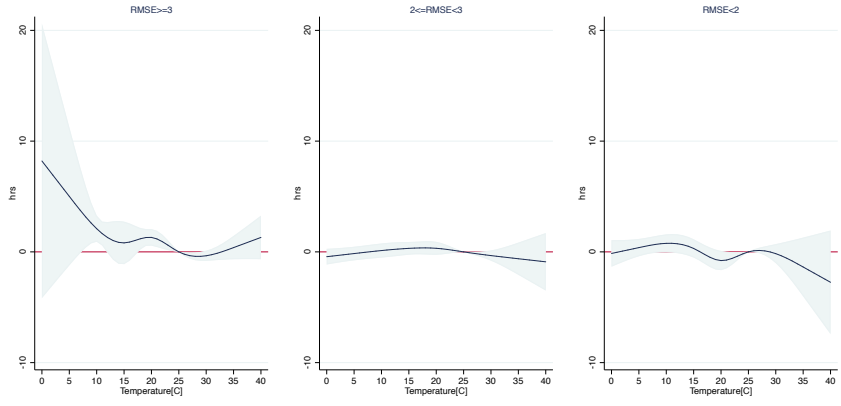
Note: Dependent variable is the half-year rolling $RMSE$ for daily $Tmax$ forecasts; Standard errors in parentheses clustered by cities; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Factors indicate whether the regression include only economic or both economic and physical factors; FE indicates whether a date fixed effect is included in the regression; Sample includes all 342 cities and 685 non-missing forecasts days of 2011 and 2015; Observations with either missing economic or physical factors are excluded; OOS R^2 indicates the average across 10-fold cross-validation out-of-sample R^2.

Table 1.C.4: RMSE Variation Decomposition Regression with Physical Controls, Full Sample 2011 and 2015

	(1)	(2)	(3)
Water Resources ($10^8 m^3$)	-0.001	-0.001	-0.002
	(0.001)	(0.001)	(0.001)
Greenland Area [$10^4 km^2$]	-4.068**	4.795	3.254
	(1.860)	(3.167)	(3.385)
Elevation [km]	1.066***	0.935***	0.900***
	(0.184)	(0.220)	(0.212)
No. Weather Stations	-0.000	0.028	0.027
	(0.040)	(0.038)	(0.039)
Area of Lakes [km^2]	-0.562	-0.374	-0.652
	(0.864)	(1.021)	(0.998)
Lenghth of Rivers [Deg]	0.006***	0.003	0.003
	(0.002)	(0.004)	(0.004)
Annual Avg. $Tmax$ [C]	-0.038	-0.057*	-0.070**
	(0.031)	(0.032)	(0.033)
Annual STD. $Tmax$ [C]	-0.117*	-0.145**	-0.167***
	(0.061)	(0.060)	(0.063)
Annual Avg. Precip. [mm]	-0.007	-0.017	0.014
	(0.039)	(0.041)	(0.042)
Factors	Physical	Both	Both
Date FE	No	No	Yes
Observations	156759	156759	156759
Adjusted R^2	0.456	0.493	0.521
OOS R^2	0.456	0.493	0.519

Note: Dependent variable is the half-year rolling $RMSE$ for daily $Tmax$ forecasts; Annual averages for $Tmax$ or Precip. are average real weathers of the current year; Standard errors in parentheses clustered by cities; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Factors indicate whether the regression include only physical or both economic and physical factors; FE indicates whether a date fixed effect is included in the regression; Sample includes all 342 cities and 685 non-missing forecasts days of 2011 and 2015; Observations with either missing economic or physical factors are excluded; OOS R^2 indicates the average across 10-fold cross-validation out-of-sample R^2.

Figure 1.B.8: Non-Parametric Interactive Regression



Note: Left to Right, Bins of RMSE in Decreasing Order $[3C, \infty)$, $[2C, 3C)$, $[1C, 2C)$.

1.D Additional Plots for Results and Robustness Checks

Figure 1.D.1: Simple Regression with Realized $Tmax$

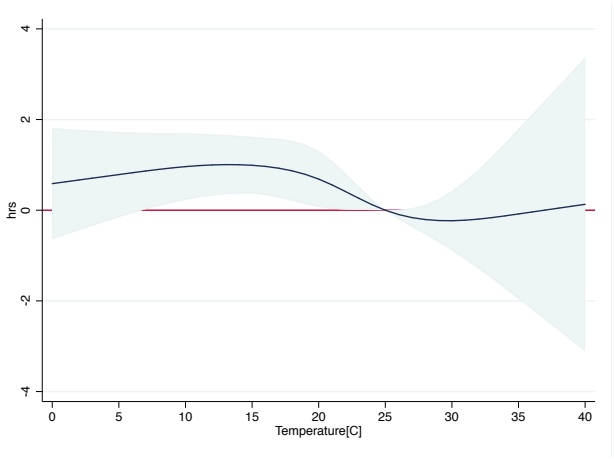
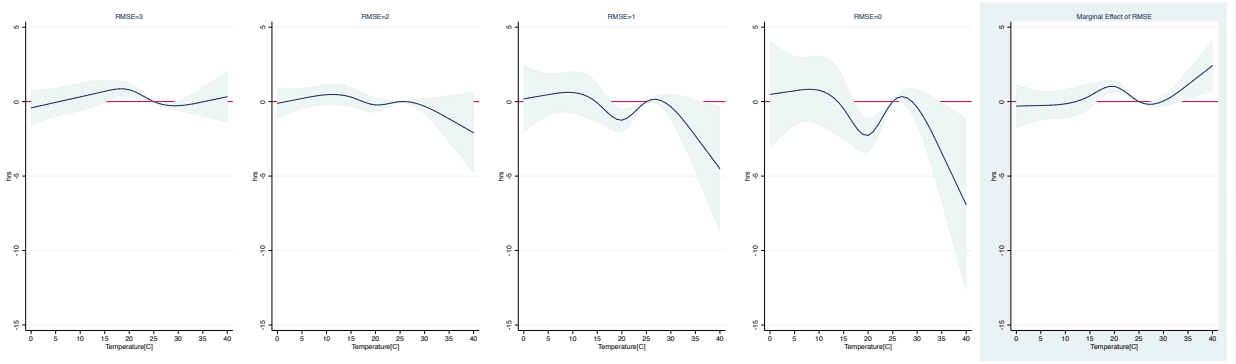


Table 1.D.1: Interactive Regression Table with Five Cities - Labor Response to Forecast Temperatures Relative to 25C

$Tmax^{forecast}$ [C]	Jinan	Changchun	Beijing	Kunming	Chengdu
Avg. $RMSE$ [C]	1.471	1.891	2.494	2.978	3.773
0	0.042	-0.084	-0.265	-0.410	-0.648
	(0.816)	(0.563)	(0.403)	(0.576)	(1.086)
5	0.346	0.237	0.081	-0.045	-0.251
	(0.522)	(0.383)	(0.349)	(0.484)	(0.823)
10	0.543	0.481	0.391	0.320	0.202
	(0.520)	(0.380)	(0.344)	(0.478)	(0.815)
15	0.097	0.260	0.493*	0.681*	0.989
	(0.493)	(0.367)	(0.292)	(0.369)	(0.628)
20	-0.763**	-0.334	0.282	0.776***	1.589***
	(0.321)	(0.267)	(0.232)	(0.251)	(0.352)
25	0.000	-0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)	(.)
30	-0.331	-0.315	-0.293	-0.275	-0.246
	(0.360)	(0.292)	(0.209)	(0.171)	(0.206)
35	-1.758*	-1.280	-0.594	-0.042	0.863
	(1.029)	(0.847)	(0.619)	(0.496)	(0.526)
40	-3.368*	-2.352	-0.896	0.275	2.198**
	(1.778)	(1.469)	(1.085)	(0.886)	(0.950)

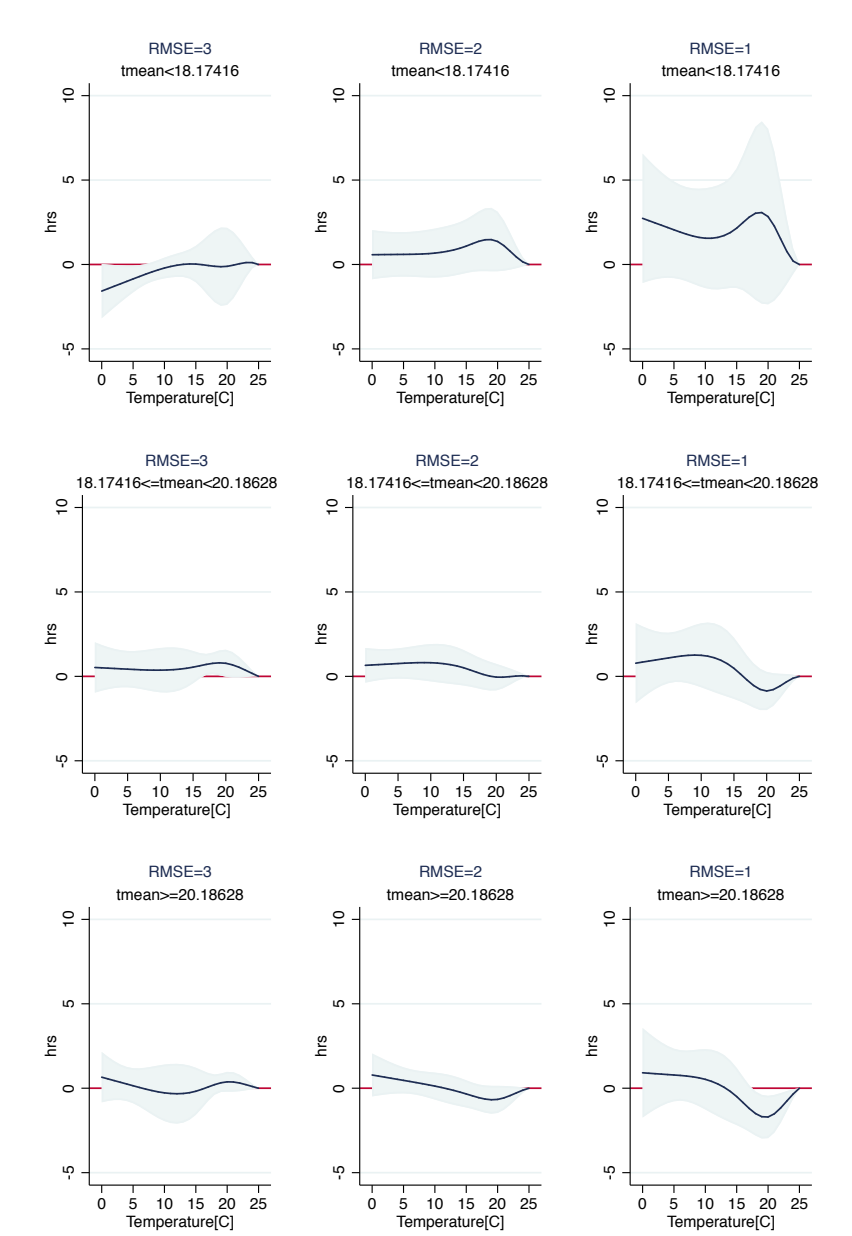
Note: Average $RMSE$ is average rolling $RMSE$ of half a year in year 2011 and 2015 with non-missing forecasts (685 days) for the corresponding city.

Figure 1.D.2: Main Interactive Regression Extrapolated to Perfect Forecasts



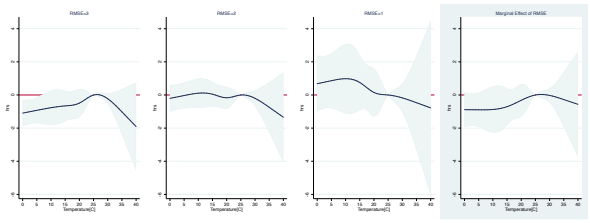
Note: Left to right, RMSE decreases 3,2,1,0, marginal effect of RMSE.

Figure 1.D.3: Interactive Regression with Tercile Separation by Current Year Average Real Tmax

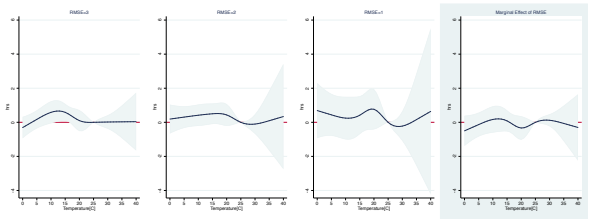


Left to Right, RMSE decreases 3,2,1; Top to Bottom, Cold, Medium, Hot.

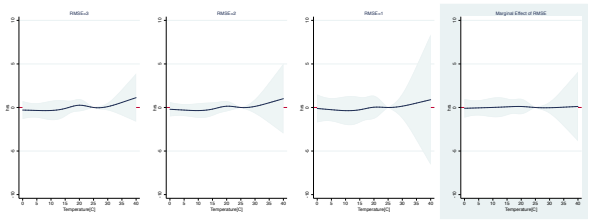
Figure 1.D.4: Interactive Regression Subsample Analysis by CHNS Primary Occupations, Part I



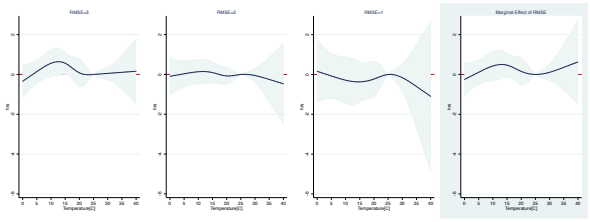
(a) Senior Professionals/Technicals



(b) Junior Professionals/Technicals

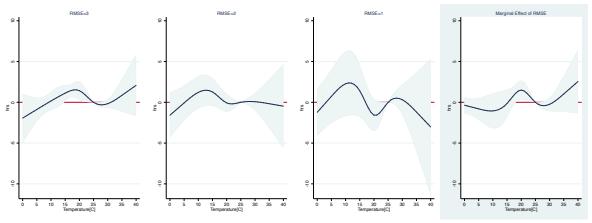


(c) Administrators/Executives/Managers

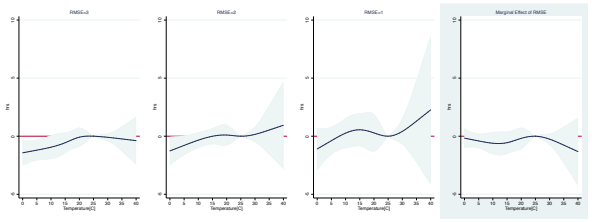


(d) Office Staffs

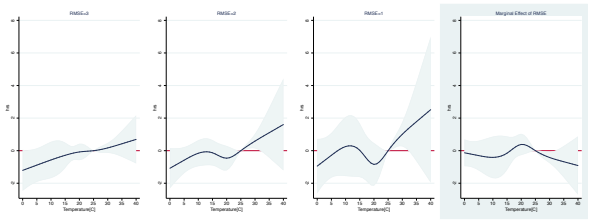
Figure 1.D.5: Interactive Regression Subsample Analysis by CHNS Primary Occupations, Part II



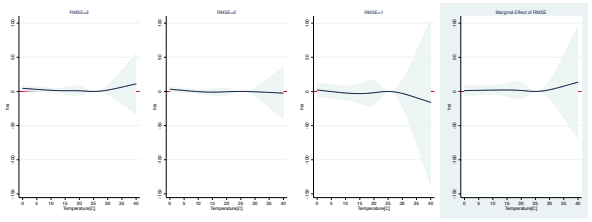
(a) Farmers/Fishers/Hunters



(b) Skilled Workers

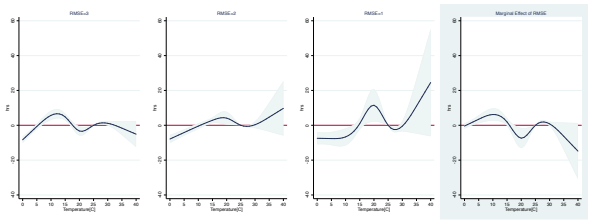


(c) Non-Skilled Workers

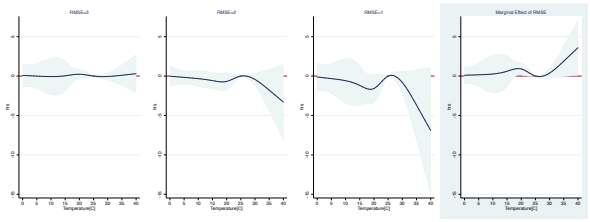


(d) Army Officers/Police Officers

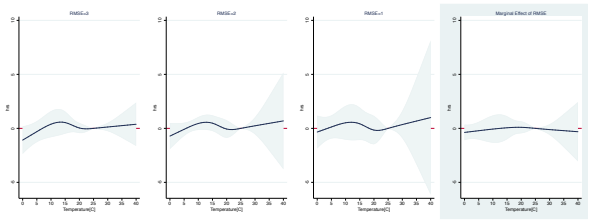
Figure 1.D.6: Interactive Regression Subsample Analysis by CHNS Primary Occupations, Part III



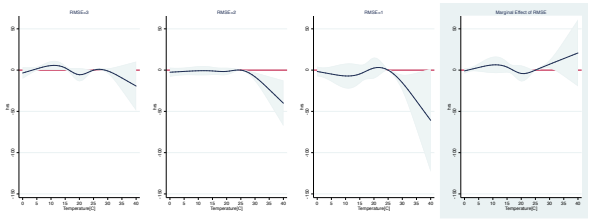
(a) Soldiers/Policemen



(b) Drivers

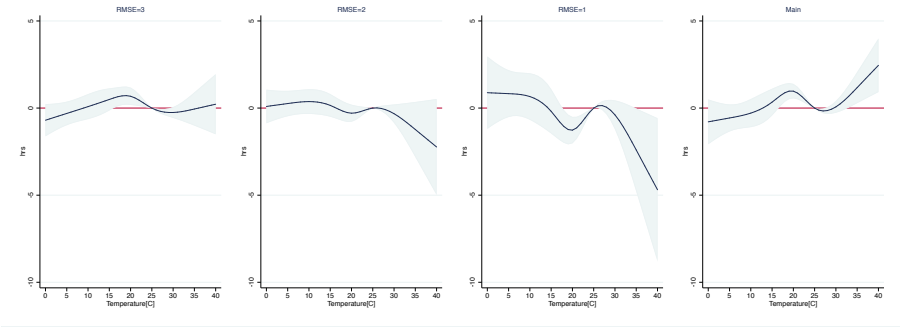


(c) Service Workers

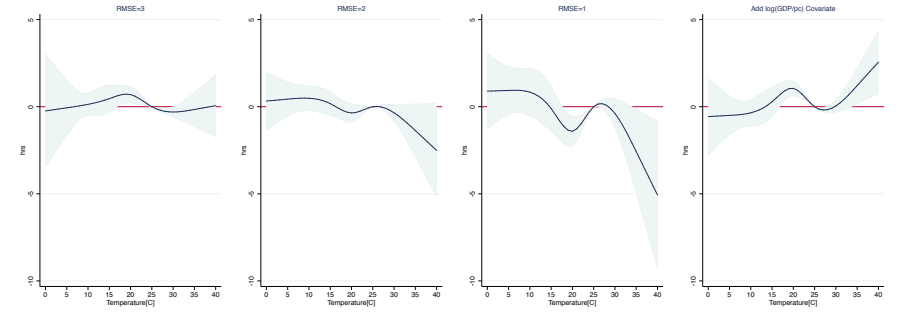


(d) Athletes/Actors/Musicians

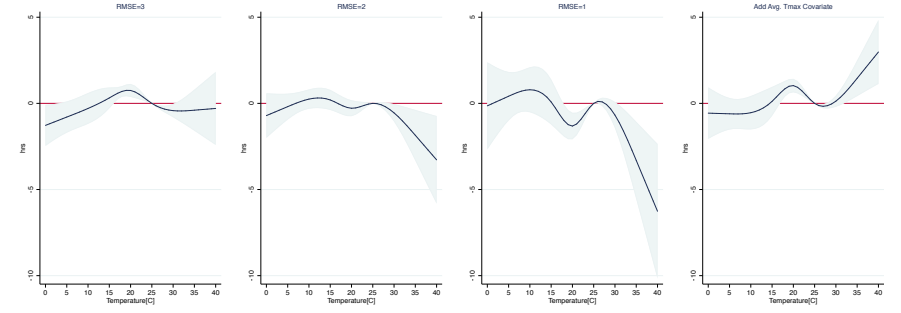
Figure 1.D.7: Double and Triple Interactive Regression



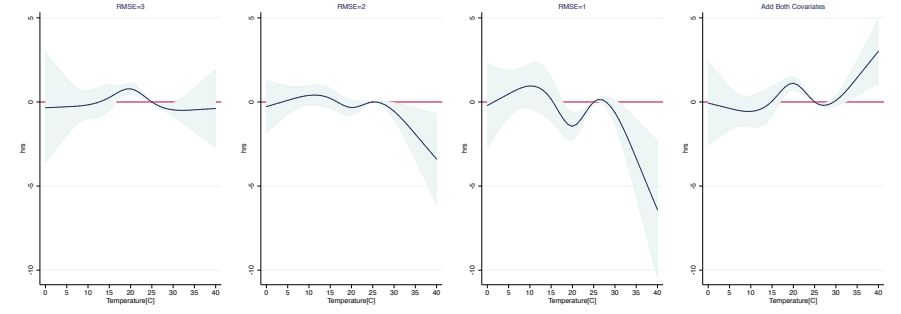
(a) Original Single Interactive



(b) Double Interactive with Income Covariate



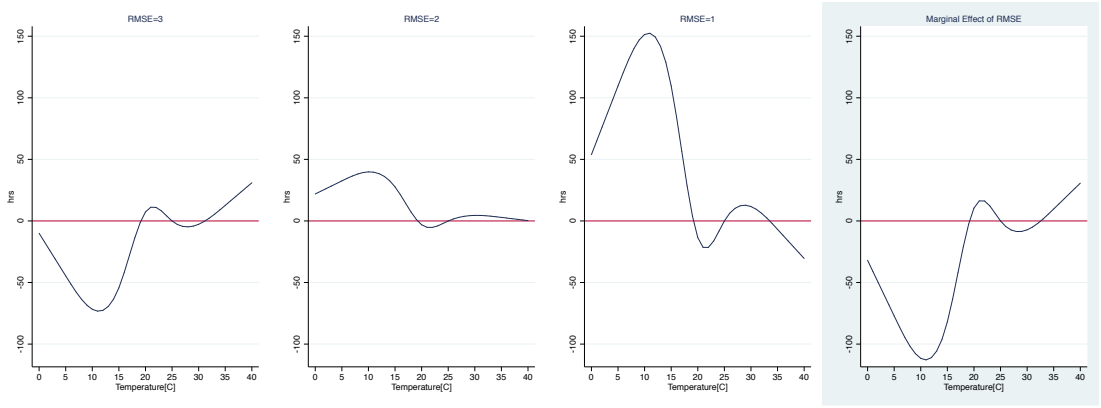
(c) Double Interactive with Climate Covariate



(d) Triple Interactive with Both Income and Climate Covariates

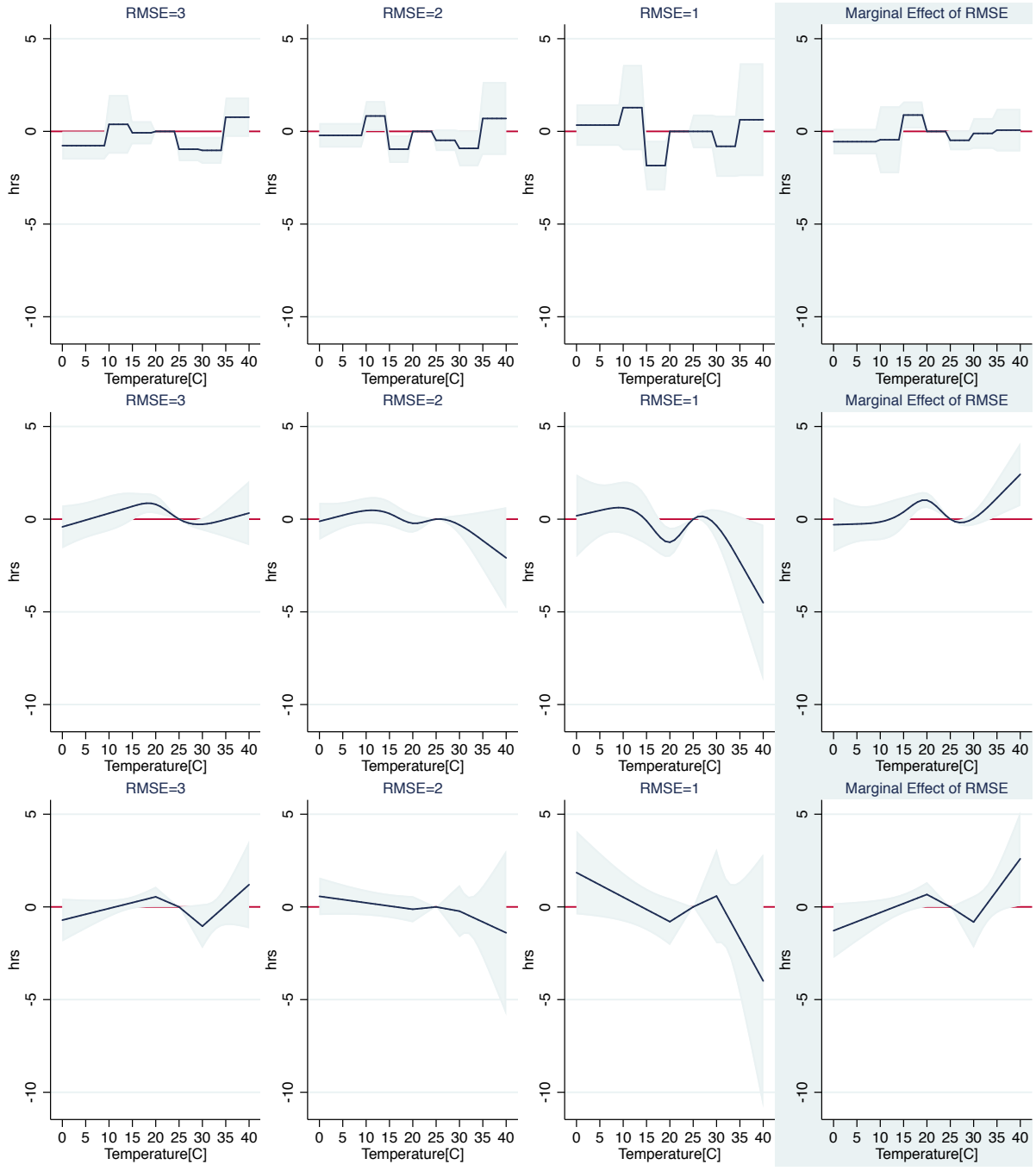
Note: Left to Right: RMSE decreases 3,2,1, Marginal effect of RMSE covariate; Array estimates performed fixing sample average income or climate covariates.

Figure 1.D.8: Instrumental Variable Interactive Regression



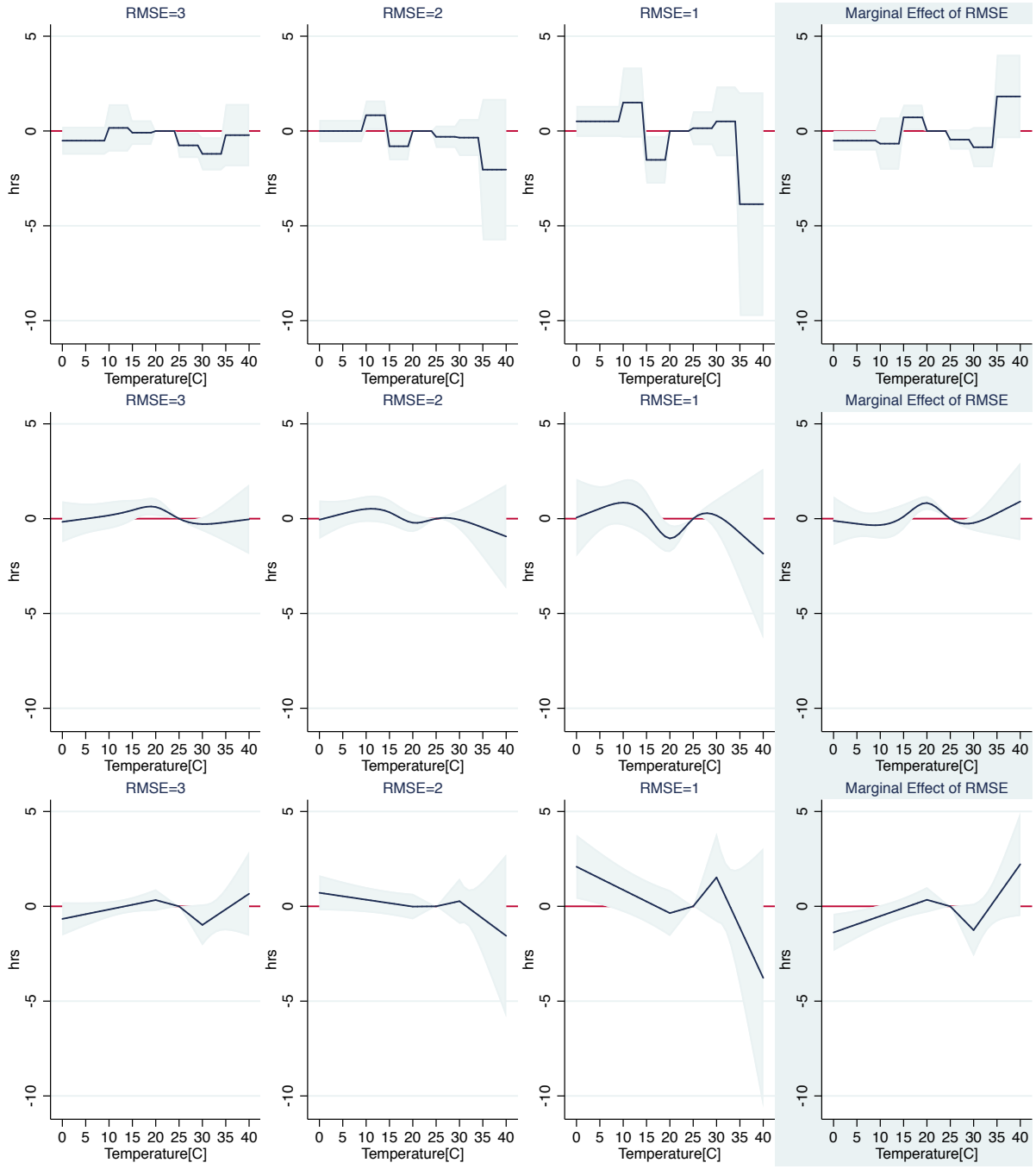
Note: IV being elevation; Left to right, RMSE decreases 3,2,1, marginal effect of RMSE; Confidence intervals are wide and dropped for display purpose.

Figure 1.D.9: Interactive Regression with Half-Year Rolling Window $RMSE$



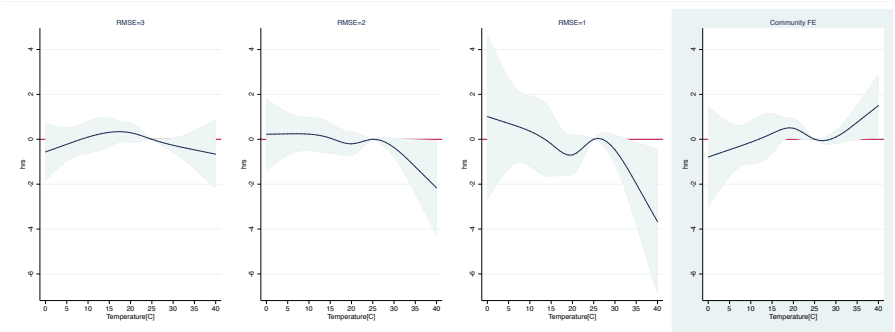
Note: Left to Right, RMSE decreases 3,2,1; Top to Bottom, Bins, Restricted Cubic Spline, Linear Spline.

Figure 1.D.10: Interactive Regression with Four-Month Rolling Window $RMSE$

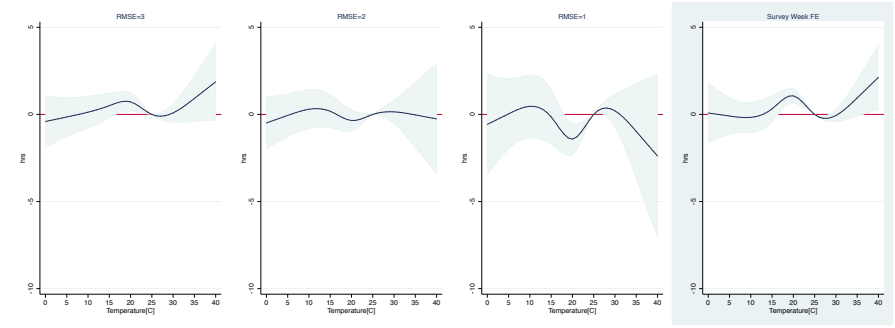


Note: Left to Right, RMSE decreases 3,2,1; Top to Bottom, Bins, Restricted Cubic Spline, Linear Spline.

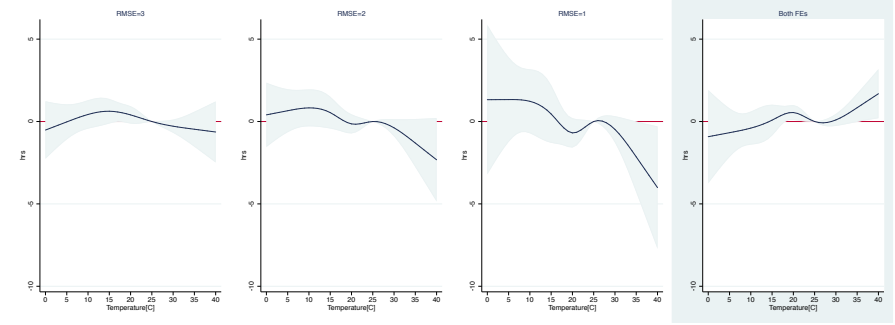
Figure 1.D.11: Interactive Regression with Different Fixed Effects Settings



(a) Community FE + Month FE



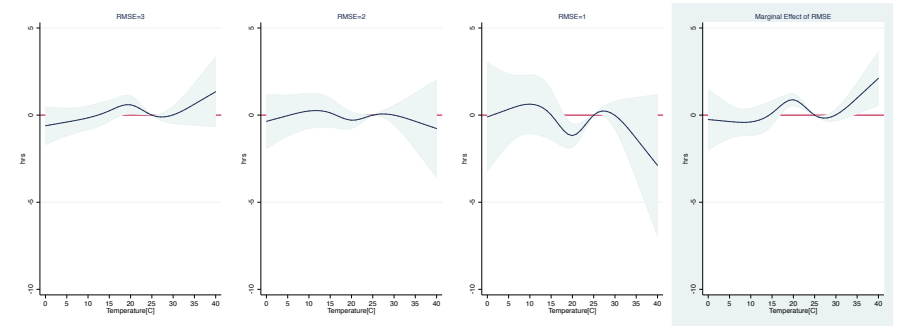
(b) City FE + Week FE



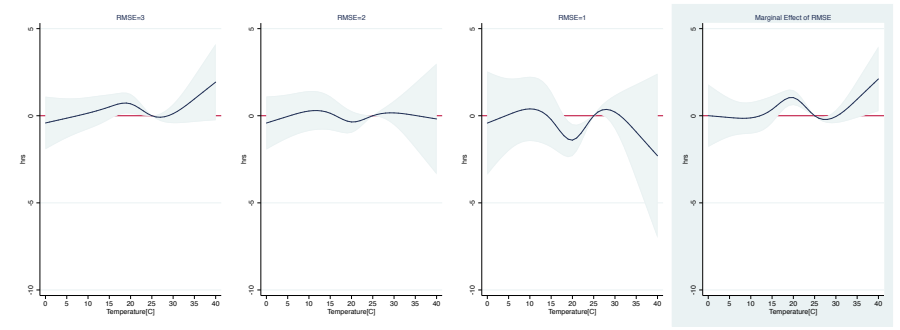
(c) Community FE + Week FE

Note: Left to Right, RMSE decreases 3,2,1, marginal effect of RMSE.

Figure 1.D.12: Interactive Regression on Restricted Sample on Weekly Working Hours



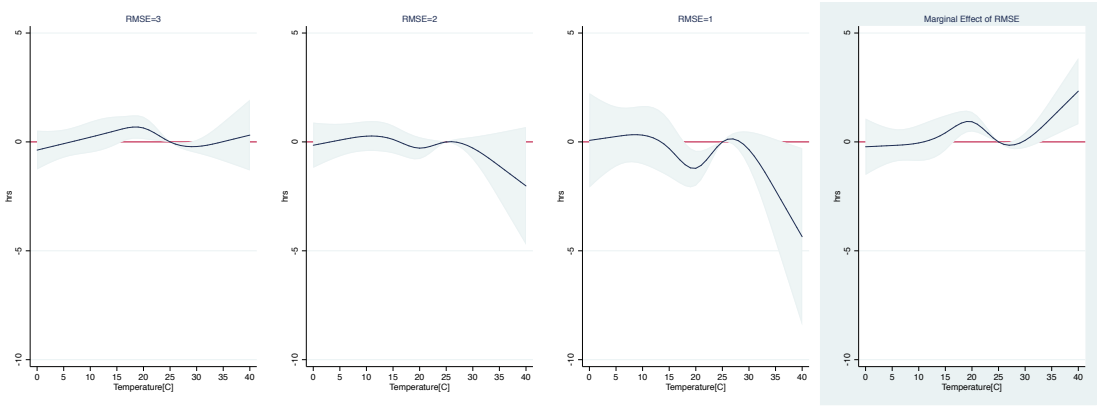
(a) Trim Sample to in $(0, 84]$



(b) One-Side Winsorization to 84 Hours Max

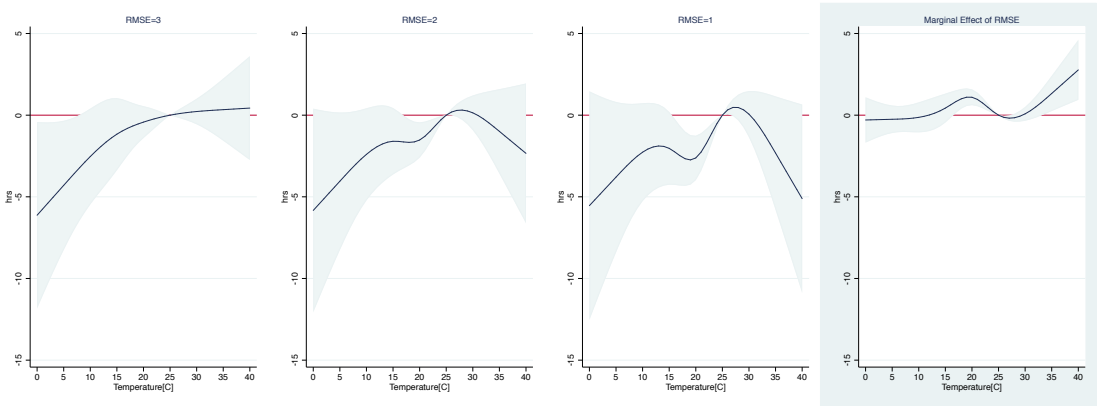
Note: Left to right, RMSE decreases 3,2,1, marginal effect of RMSE.

Figure 1.D.13: Interactive Regression with Extra Economic and Demographic Controls



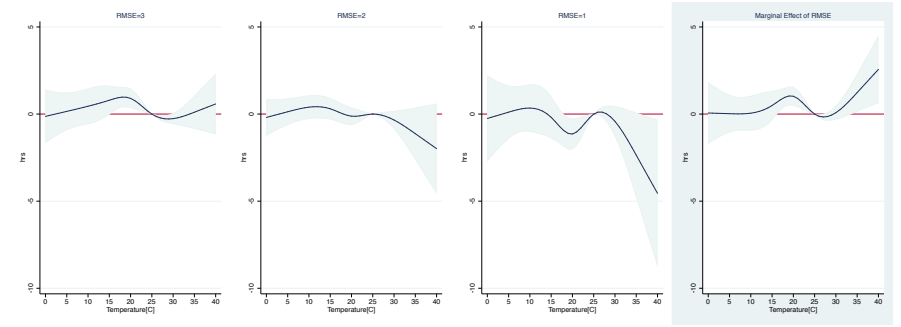
Note: Controls added include log GDP per capita, log population, age, gender and college degree dummy; Left to right, RMSE decreases 3,2,1, marginal effect of RMSE.

Figure 1.D.14: Interactive Regression with Realized Temperature Splines Controlled

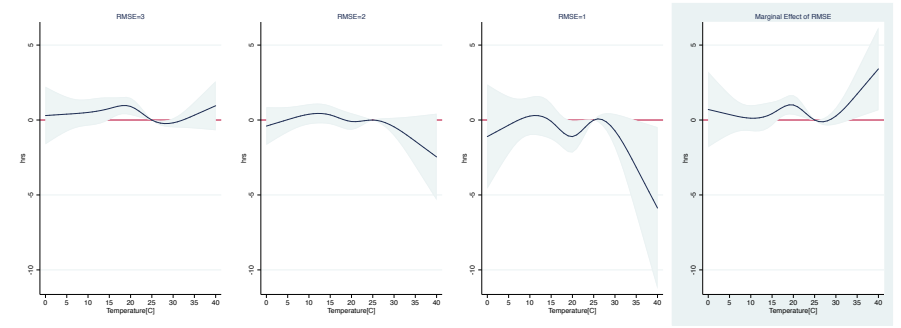


Note: Including same 5 knots restricted cubic spline of realized $Tmax$ as linear controls; Left to right, RMSE decreases 3,2,1, marginal effect of RMSE.

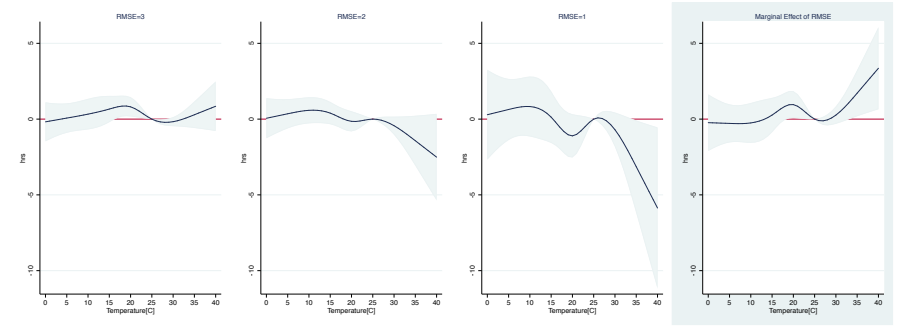
Figure 1.D.15: Interactive Regression with $RMSE$ Defined with Smaller Temporal Variations



(a) $RMSE$ of Current Year



(b) $RMSE$ of Previous Year



(c) $RMSE$ of 2010

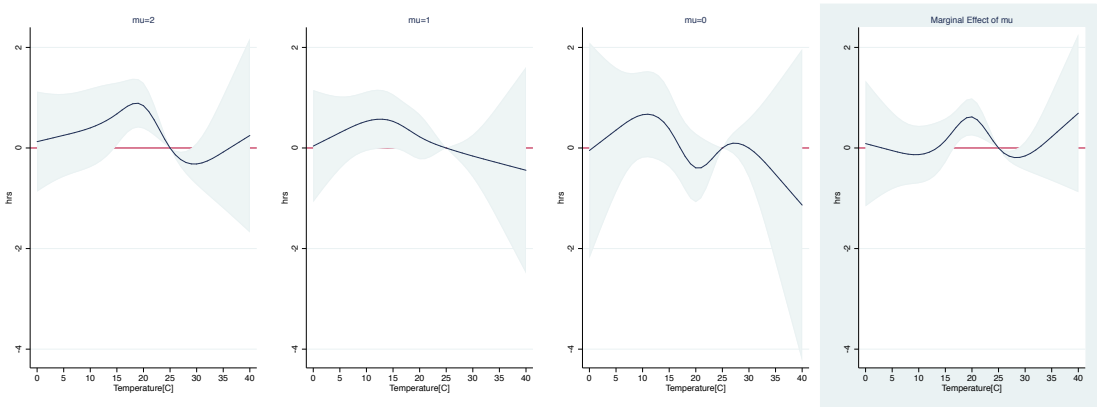
Note: Left to right, RMSE decreases 3,2,1, marginal effect of RMSE.

Table 1.D.2: Summary Statistics Comparing RMSE with Alternative Forecast Accuracy Metrics

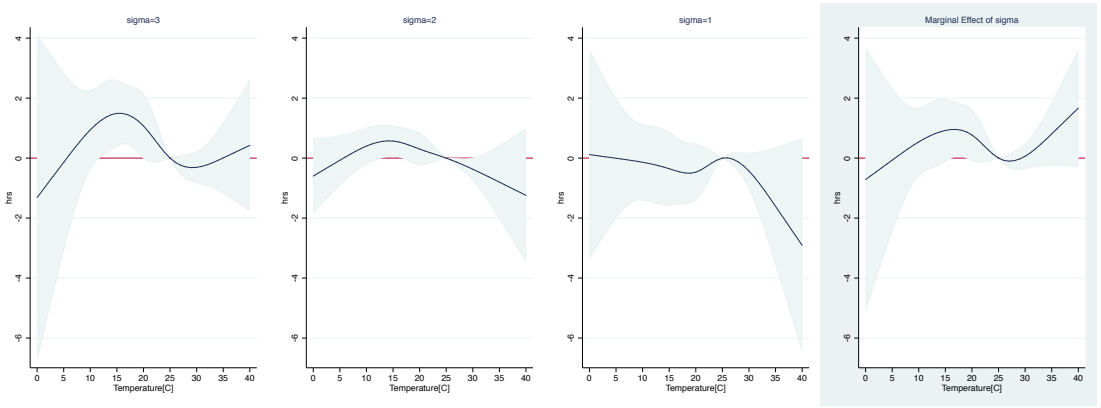
	Labor Sample					All Cities All Days 2011 and 2015				
Metric	N	Mean	STD	Min	Max	N	Mean	STD	Min	Max
Main RMSE	2968	2.478	0.676	1.351	4.306	234270	3.323	1.837	0.991	13.688
Abs. Mean Error	2968	1.091	0.847	0.005	3.257	234270	1.935	2.150	0.000	13.378
Error STD.	2968	2.100	0.554	1.227	4.032	234270	2.371	0.684	0.947	5.938
Rational RMSE	2968	1.990	0.480	1.227	3.729	233586	2.265	0.622	0.722	6.139
AR RMSE	2968	2.880	0.535	1.096	3.953	233586	2.932	0.678	0.367	4.816
Hot RMSE	2968	1.009	1.349	0.000	4.354	234270	1.316	2.005	0.000	20.710
Cold RMSE	2968	1.328	1.145	0.000	4.247	234270	1.937	2.208	0.000	13.680
Max Abs. Error in Half Year	2968	7.554	2.219	4.025	15.531	234270	9.283	3.245	2.753	24.865
Max Abs. Error in One Month	2968	5.030	1.754	1.867	12.111	233928	6.789	3.018	0.000	24.865

Note: Rational and AR generated $RMSE$ has fewer observations in the all city all dates sample because they are both run on rolling window of 183 days, hence the earliest days of 2011 could have missing $RMSE$ because not all previous rationalized (or AR predicted) forecasts can be estimated (my sample starts from 2010). For maximum absolute error over a month, more observations are missing from the early 2011 sample because there are more than a month missing data for early that year.

Figure 1.D.16: Interactive Regression with RMSE Breakdown Components



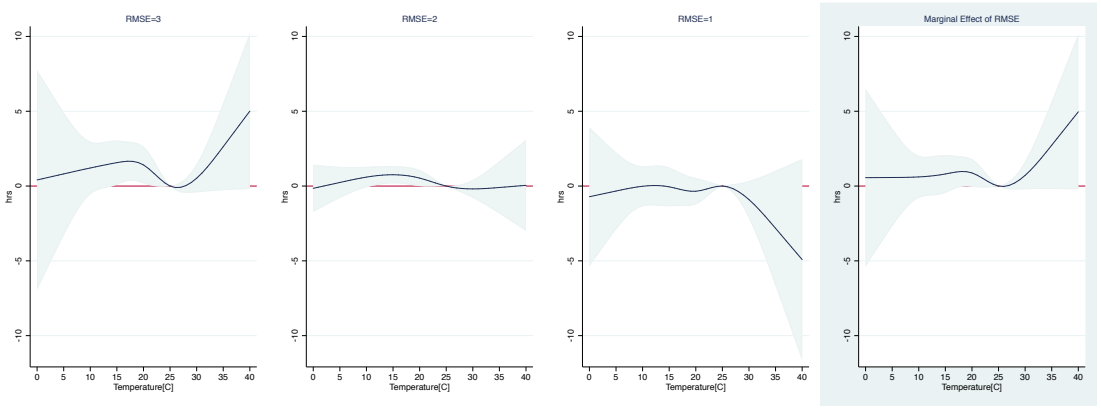
(a) Covariate Abs. Error Mean



(b) Covariate Error STD

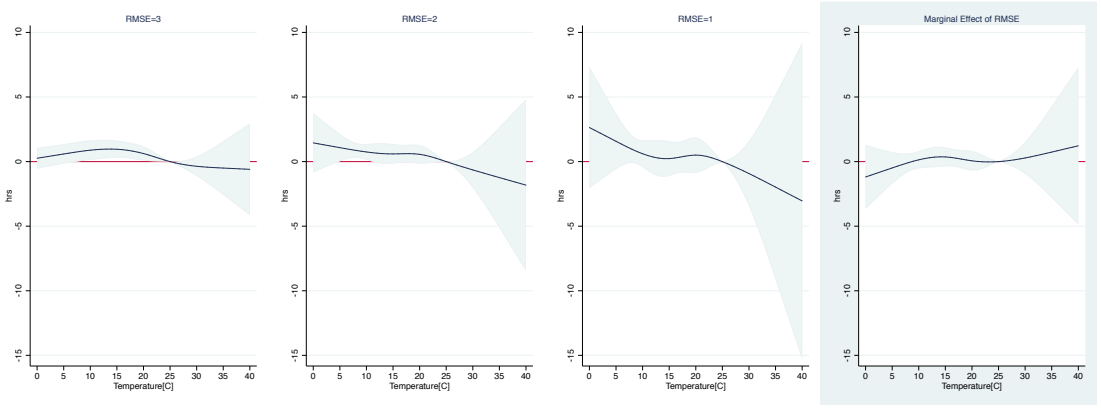
Note: Absolute mean errors and standard deviation on rolling window of 183 days (half a year); Left to Right, Global, RMSE decreases 3,2,1, Marginal effect of covariate.

Figure 1.D.17: Interactive Regression with Rationalized Forecasts



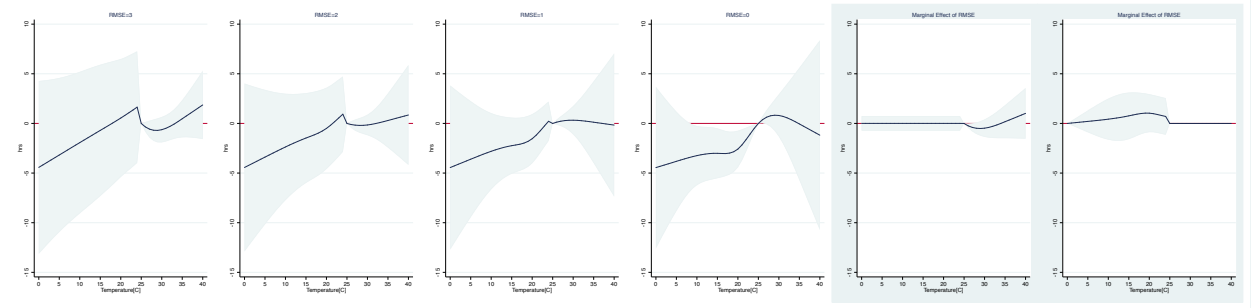
Note: Rationalization runs with ordinary least squared regression of $Tmax^{real}$ on $Tmax^{forecast}$ on rolling window of 183 days for each city each date, estimating $Tmax^{rational} = \hat{\alpha} + \hat{\beta} Tmax^{forecast}$; RMSE kept at rolling window of 183 days (half a year); Left to right, RMSE decreases 3,2,1, marginal effect of RMSE.

Figure 1.D.18: Interactive Regression with Auto-regression Predicted Forecasts



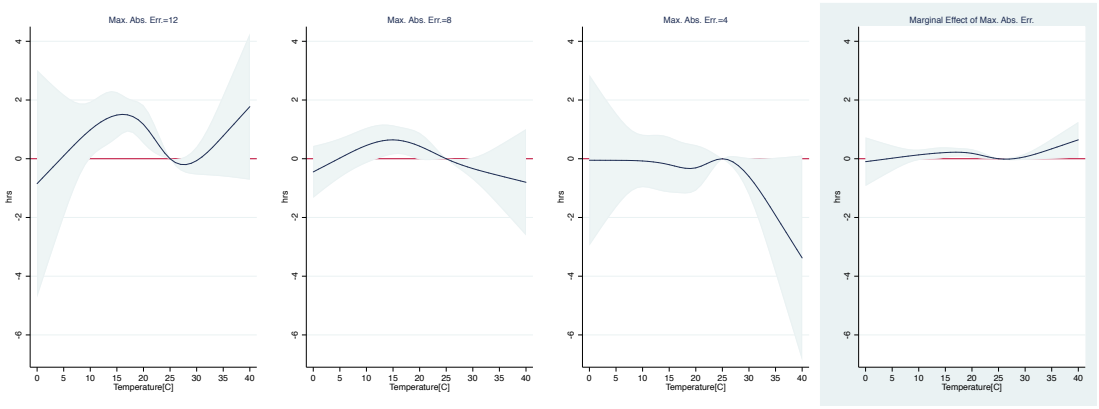
Note: AR(7) runs with ordinary least squared regression of real temperature lags on rolling window of 183 days for each city each date, estimating $Tmax^{AR} = \hat{\alpha} + \sum_{k=1}^{7} \hat{\beta}_k Tmax^{real}_{-k}$; RMSE kept at rolling window of 183 days (half a year); Left to right, RMSE decreases 3,2,1, marginal effect of RMSE.

Figure 1.D.19: Interactive Regression with Splitted RMSE Interactions

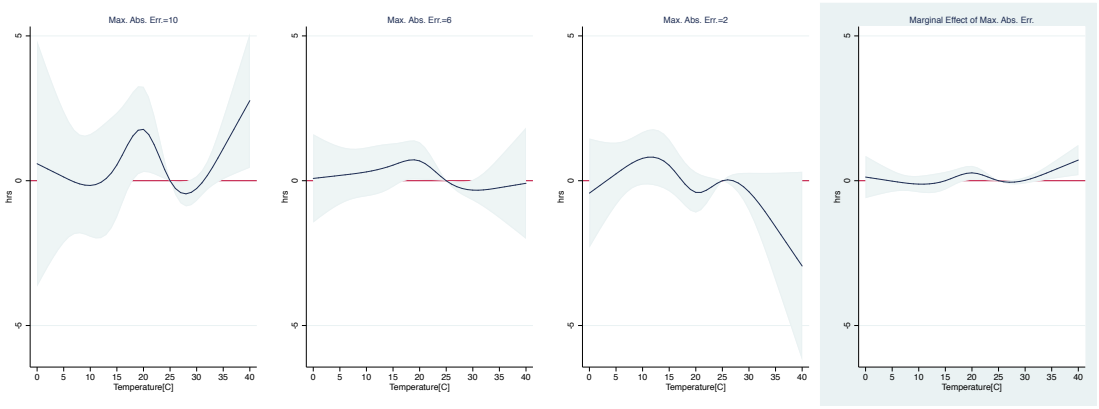


Note: For $Tmax^{forecast} < 25C$ interact with RMSE for forecasts only below $25C$ (zero otherwise); For $Tmax^{forecast} \geq 25C$ interact with RMSE for forecasts only greater or equal $25C$ (zero otherwise); Both RMSE adapted on rolling window of 183 days (half a year); Left to right, RMSE decreases 3,2,1, marginal effect of hot end RMSE, marginal effect of cold end RMSE.

Figure 1.D.20: Interactive Regression with Maximum Absolute Error



(a) Maximum Absolute Error over Half-Year



(b) Maximum Absolute Error over One-Month

Note: Absolute errors taken maximum over rolling window of 183 days (half a year) or 30 days (one month); Left to Right, maximum absolute error decreases 12,8,4, Marginal effect of covariate.

1.E Detailed Description of Valuation Steps

This section serves as an amendment to Section 5.2, in description of the detailed steps for valuation to estimate the single-period utility maximization model with the baseline regression estimates.

Step 1: Referencing α

In my model, the scaler parameter α is the constant return of labor. This quantity is relate to the marginal labor return to wage rate, $\frac{\partial l^*}{\partial p_l} = -\frac{1}{2\alpha}$. Further related to the more studied quantity, the elasticity of labor supply η, I approximate this parameter with $\alpha \approx -\frac{1}{2}(\frac{\Delta l}{\Delta p_l})^{-1} \approx -\frac{1}{2}(\eta\frac{l}{p_l})^{-1}$. The numeric variables I use are:

- $\eta = 0.353$: I directly reference the elasticity of labor supply η in China from Li (2016). This paper estimates the elasticity quantity with choice-based conjoint analysis on the China Urban Labor Survey (CULS) with six cities (sample worker age 16-64, almost the same as my sample). I take its estimate for all workers in 2010.

- l = 7.496 hours, p_l = 20.548 2015 Yuan: In my 2011 and 2015 CHNS labor sample, I summarize the average wage per hour and average hours worked per day from the variables of previous year wage income and working time.

I estimate the parameter $\alpha = -3.882$. This could be a relatively conservative estimate, as elasticity of labor supply may decrease over time as the country get more industrialized[73].

[73] In this reference paper, the estimate for 2001 is much larger than that of 2010.

In the main section, sensitivity analysis with alternative range of α is performed.

Step 2: Identifying $l^* = g(f, a)$ and $\beta(w)$

As argued above, $\beta(w) = -2\alpha g(w, a^*)$ can be estimated by the extrapolated labor response under $RMSE = a^*$. To compute the utility function, I also need to estimate the optimal labor function $l^* = g(f, a)$. Therefore, I can directly use the representation from the baseline interactive regression Figure 1.3 with $a^* = 1$:

- $l^* = g(f, a) = \bar{l} + [f(f; \beta_0) - f(25; \beta_0)] + a \times [f(f; \beta_1) - f(25; \beta_1)]$: With $\bar{l} = 7.410$ hours is the labor under reference temperature $f = 25$ summarized from the sample[74], I back up $l^* = g(f, a)$ from the labor response relative to $25C$ estimated in Figure 1.3 Panel (c).

- $\beta(w) = -2\alpha(\bar{l} + [f(w; \beta_0) - f(25; \beta_0)]) + a^* \times [f(f; \beta_1) - f(25; \beta_1)]$: With the extrapolated labor function under perfect forecasts and α parameter referenced, the non-linear function $\beta(w) = -2\alpha g(w, a^*)$ is estimated.

Here the components of the non-linear labor response functions $f(.; \beta_0), f(.; \beta_1)$ (for main analysis, they are restricted cubic splines) have coefficient vectors β_0, β_1 directly estimated from the baseline interactive regression. To account for uncertainties, I also sample β_0, β_1

[74]Assuming that the equilibrium labor $\bar{l}$ at $Tmax^{forecast} = 25C$ is roughly invariant with the forecast accuracy metric $RMSE$ ($RMSE$ as a linear control in this regression does have its coefficient statistically insignificant from zero), I estimate $\bar{l}$ as the 7-day average labor hours within the city-week in my regression labor sample where its average $Tmax^{forecast}$ is closest to $25C$ (the value is $25.022C$). In this estimation, I assume fixing the other linear controls $\mathbf{X}$. One can regard this as the case of precipitation zero (the in-sample average daily precipitation is about 3mm).

with Monte-Carlo runs (300 trials) drawing from the multinomial distribution with the regression estimates as mean and their variance-covariance matrix.

Step 3: Simulating $\bar{V}(w,a)$

To calculate $\bar{V}(w,a)$, I average the realized utility $u(l^*,w)$ for any pair of (w,a) across simulations of $f|w,a$. In real data, this conditional distribution is approximately normal[75]. Therefore, I simulate (with 300 draws) from the normal distribution $f|w,a \sim N(w-\mu,\sigma^2)$ with selections of two parameters, the error $(w-f)$ mean μ and standard deviation σ:

- $\mu = 0$: For main valuation analysis, I keep a constant average bias parameter $\mu = 0$, which is the sample minimum (see Appendix 1.H.2).
- $\sigma = 1, 1.25, ..., 4$: I allow the standard deviation of forecasts distribution to vary in a range that covers from 1 to close to the maximum of the half-year rolling $Tmax$ forecasts error standard deviation in my labor sample.

Note that these assumptions are restricted to cases of unbiased forecasts with $\mu = 0$ with positive uncertainties $\sigma > 0$ in order to simplify computations. These assumptions are consistent with the forecasts sample observations, where μ can get close to zero but σ is never below 0.9. My analysis assume labors most care about the $RMSE$ instead of the two components, μ, σ so these simplifications are reasonable. In main section of sensitivity analysis, the estimates on actual μ, σ pairs observed in the dataset are shown with the simplification

[75]Kolmogorov-Smirnov tests prove that 98.3% of the days in my labor sample does not reject the normality test at 5%.

results and they match.

$\bar{V}(w,a)$ is then estimated for the series of real temperatures $w = 0,1,...,40$ under the $RMSE$ metric simulated at $a \in [1,4]$ by $RMSE = a \approx \sqrt{\mu^2 + \sigma^2}$. By the quadratic utility design, $\bar{V}(w,a)$ achieve maximum under any given w if and only if $a = a^*$ ($\sigma = a^*$), which I also compute as a reference for later use.

Step 4: Aggregating to $V(a)$

For main valuation analysis, I estimate $V(a)$ by taking the average of simulated $\bar{V}(w,a)$ weighted by a selected real temperature distribution $p_0(w)$. To keep it constant, I maintain $p_0(w)$ as the empirical distribution of $Tmax^{real}$ over the pooled sample including all 342 cities and all 730 days in 2011 and 2015. I compute $V(a)$ for series of $a \in [1,4]$ as described with $\bar{V}(w,a)$. For display purpose, I normalize $V(a)$ subtracting the perfect forecast case $a = a^* = 1$ and inflate the values by 365 days of a year (because my labor sample does not exclude weekends and holidays). In interpretation, the final value would be the estimated value loss per worker per year given forecast $RMSE = a$ relative to the perfect forecasts case with $RMSE = 1$. The same process is repeated for 300 Monte-Carlo draws described with the estimation of $l^* = g(f,a)$ and $\beta(w)$ to generate confidence intervals.

1.F Additional Plots for Valuation and Sensitivity Analysis

Figure 1.F.1: $V(a)$ Relative to $V(1)$, Per Labor Per Year, Different α.

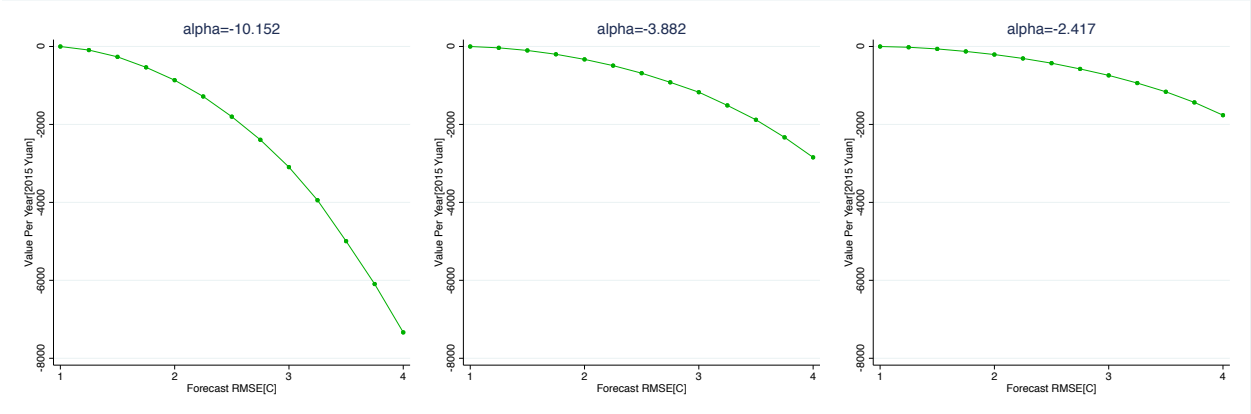
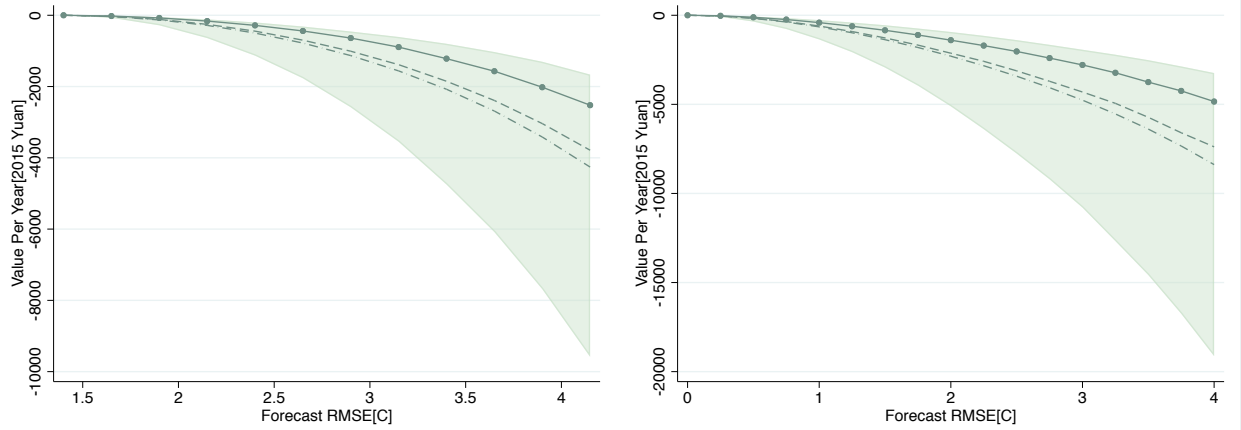


Figure 1.F.2: $V(a)$ Relative to $V(a^*)$, Per Labor Per Year, For Different a^*

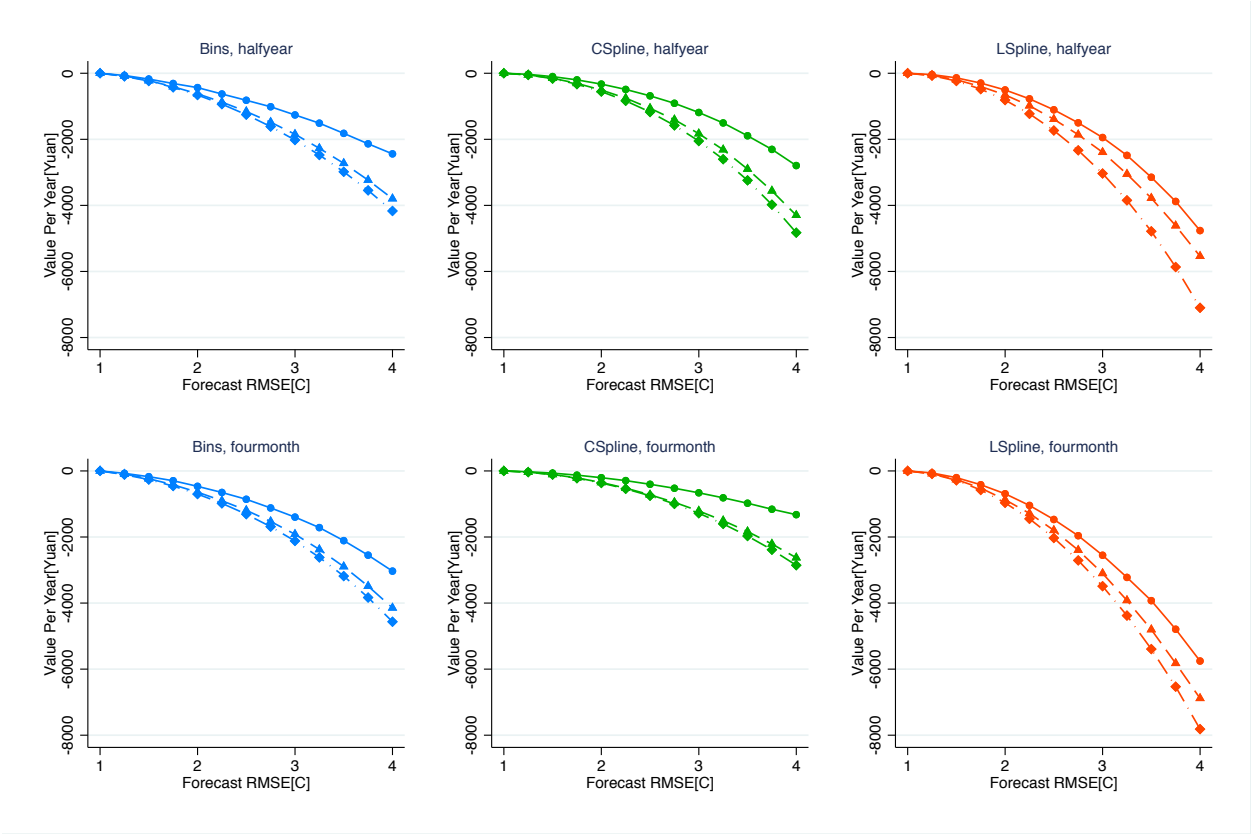


(a) $a^* = 1.4$, labor sample minimum

(b) $a^* = 0$, perfect forecasts

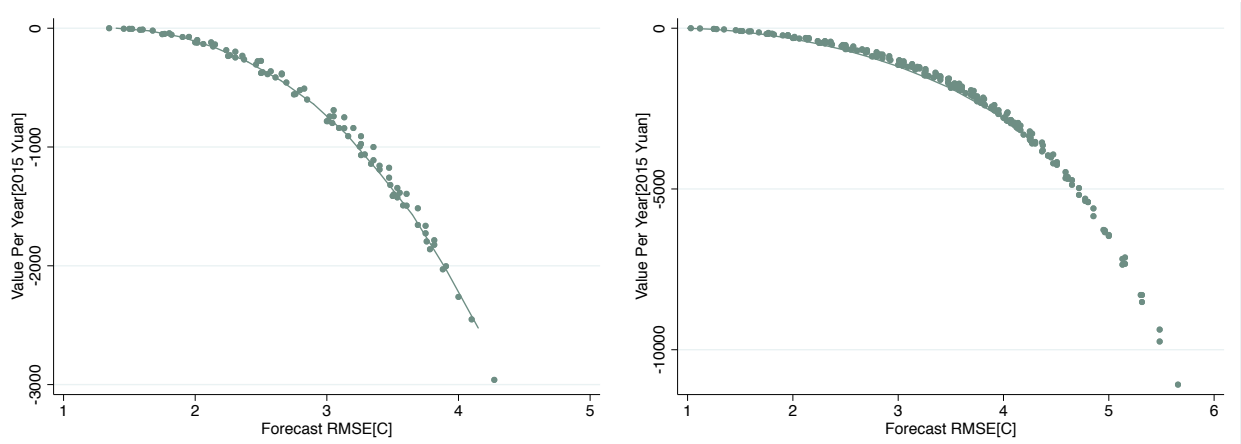
Note: MC generated 95% confidence interval in shade, dashed line MC mean, dash-dot line MC median.

Figure 1.F.3: $V(a)$ Relative to $V(1)$, Per Labor Per Year, Different Specifications and Different $RMSE$ Rolling Window



Note: Solid Line/Circle, direct estimate, Dashed Line/Triangle, MC median, Dashed-Dot Line/Diamond, MC average.

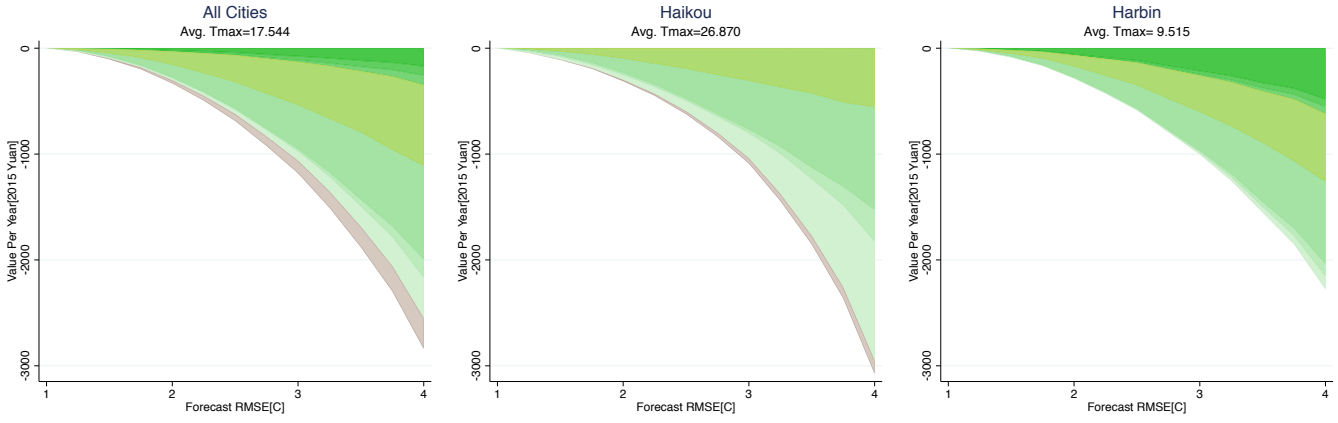
Figure 1.F.4: $V(a)$ Relative to $V(a^*)$, Per Labor Per Year, Evaluated on Real Forecasts Data Points



(a) Labor Sample (b) All Cities Sample 2011 and 2015

Note: Selected (μ, σ) pairs from forecasts data sample summarized into 0.25×0.25 grids centering at $0, 0.25, ..., 4$; a^* taken to be the minimum across the gridded $a = \sqrt{\mu^2 + \sigma^2}$; Line indicates the simulated valuations in the main section, $a^* = 1$ and $a^* = 1.4$ respectively.

Figure 1.F.5: $V(a)$ Breakdown by Real $5C$ Temperature Bins, Per Labor Per Year Relative to $V(1)$



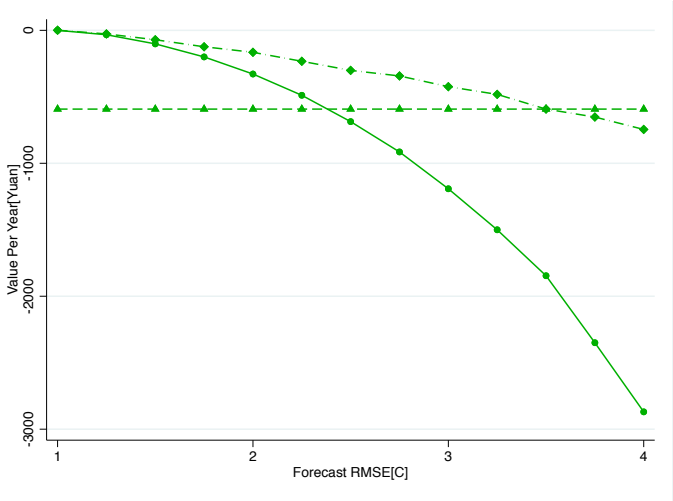
Note: Lighter color indicates higher temperature bins; Gold shades the mid temperature bin $[15C, 20C)$; Maroon shades the hottest temperature bin $[35C, \infty)$.

1.G Non-Parametric Model with Rational and Overacting Belief of Forecasts

1.G.1 Motivation and Behavior Economics Links

The reason to argue for an overreacting model of forecast belief is because my valuation using realized utilities of individual decision makers based on the regression estimated labor response do not produce the highest realized utilities (Figure 1.G.1). These potentially imply the overreaction of individuals to forecasts with not so high $RMSE$ (not so low accuracy).

Figure 1.G.1: $V(a)$ Relative to $V(1)$, Per Labor Per Year, Different Labor Response Assumptions



Note: Circle/Solid Line = full valuation (main result), Triangle/Dash Line = fixing labor response $\bar{l}$; Diamond/Dot-Dashed Line = labor varying to forecast but not to $RMSE$, always assuming $RMSE = 1$.

There are related but indiscreet descriptions for this kind of overreaction in behavior economics literatures. For example, common heuristics in real life decision makings include biased judgments under uncertainties (Twersky and Kahneman, 1974). There are also theories that can apply to decision makers being more risk averse towards bad information (Kahneman, Knetsch, and Thaler, 1991). Other theories argue that decision makers can put

different weights than actual probabilities on uncertain outcomes (Kahneman and Tversky, 1979)[76]. In a way, the general utility maximization model used in this and many other papers do not always points to what people react, but instead what they should be reacting (Thaler, 1980)[77]. These theories could explain my valuation results not being maximized over alternative labor responses less sensitive to forecasts and forecast accuracy, but overall, there are no literatures directly linking climate forecasts and behavioral models.

1.G.2 Model and Non-Parametric Estimation

With rational belief, individual decision maker shall form their expectation about the real weather shock basing on a, in this case $RMSE$, correctly as its definition and respond to the approximately normal distribution of $w|f, a \sim N(f + \mu, \sigma^2)$. Here $a = \sqrt{\mu^2 + \sigma^2}$ by definition and μ, σ are the rolling mean and standard deviation of forecast error $w - f$. However, since the empirical interactive regression results have suggested the possibility of labors overreacting to larger a as if forecasts are much more inaccurate, I would allow an overreaction factor of $\xi \geq 1$ on the standard deviation term only. Therefore, individual decision maker would form expectation about the real weather shock as:

$$w|f, a \sim N(f + \mu, \xi^2\sigma^2)$$

This would actually suggest the $RMSE$ perceived is inflated by a factor no greater than

[76]Kahneman, Daniel, and Amos Tversky. "Prospect theory: An analysis of decision under risk." In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99-127. 2013.

[77]Thaler, Richard. "Toward a positive theory of consumer choice." *Journal of economic behavior & organization* 1, no. 1 (1980): 39-60.

ξ. For $\mu = 0$ in my simulations, the inflating factor for $RMSE$ is exactly ξ. If $\xi = 1$, then there is no overreaction and the decision maker accesses the rational conditional distribution by weather forecasts.

With the normality structure of these conditional beliefs, I can then estimate this structural model with a non-parametric kernel approximation of $\beta(w)$, instead of the extrapolation method where beliefs are not assumed in the main section:

$$\beta(w) = \sum_{k=1}^{K} \beta_k(w)$$
$$\mathbb{E}[\beta(w)] = \sum_{k=1}^{K} \mathbb{E}[\beta_k(w)|f, a]$$

When β_k is a series of approximating piecewise functions on a given bin $w \in B_k$. With the specific assumptions on the functional forms of $\beta_k(w)$, $\beta(w)$ is taken to be restricted linear spline where estimates are flat after the last knot w_K:

$$\beta(w) = \beta_0 + \beta_1(w - (w - w_K)\mathbf{1}\{w > w_K\}) + \sum_{k=1}^{K-1} \gamma_k((w - w_k) - (w - w_K)\mathbf{1}\{w > w_K\})\mathbf{1}\{w > w_k\}$$

In that way, $\mathbb{E}[\beta(w)]$ can be expressed in closed form of f, a as truncated normal moments. Hence, I can directly estimate $\beta(w)$ non-parametrically with the coefficients (scaled by the same $\alpha = -3.882$) in the following regression:

$$l = b_0 + b_1(f + \xi\mu - M_K) + \sum_{k=1}^{K-1} g_k(M_k - M_K) + m'\mathbf{X} + \epsilon$$

Where with normality assumption, $f|w, a \sim N(w - \xi\mu, \xi^2\sigma^2)$ and truncated normal provides:

$$M_k = \mathbb{E}[(w - w_k)\mathbf{1}\{w > w_k\}] = (f + \mu - w_k)(1 - \Phi(\frac{w_k - \mu - f}{\xi\sigma})) + \xi\sigma\phi(\frac{w_k - \mu - f}{\xi\sigma})$$

For here, I control all the same city and month fixed effects, $RMSE$ as an independent linear control, precipitation quadratic in $\mathbf{X}$ like the baseline interactive regression. All regressors M_k are estimated on the estimated mean and standard deviation μ, σ on the same rolling window of half a year (183 days) as the $RMSE$. For this exercise, I select $K = 3$ with the temperature bin cut-offs at $w_k = 15, 20, 35$ (my explorations would give similar results for different trials of other numbers and positions of these knots).

1.G.3 Results and Valuation

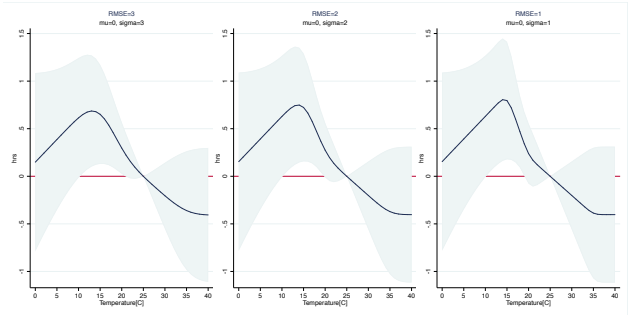
I plot the estimated labor response arrays under the main simulation setting of $\mu = 0$ and $\sigma = 1, 2, 3$, again relative to $Tmax^{forecast} = 25C$ with $\xi = 1, 2, 5, 10$ in Figure 1.G.2. Comparing with the main interactive regression, their labor decrease under hot temperature forecasts remain small both in magnitudes and in statistical significance until the overreaction factor becomes very large at $\xi = 10$. By $\xi = 10$, the labor decrease at hot end approaches the large magnitude as the main results. The medium-cold labor drop at low $RMSE$ no longer exists,

instead there is a mid-temperature labor rise. Overall, the labor response curves smooth as $RMSE$ increases, and the degree of smoothing increases with ξ.

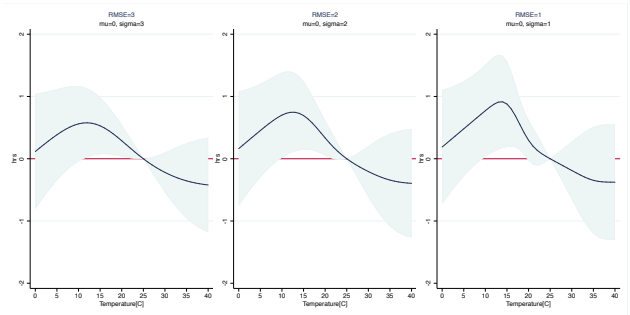
The disparity between main and non-parametric regressions suggests the possibility that $\xi \in [5, 10]$. However, the empirical determination of the optimal overreacting factor ξ gives different results. My approach is to record the residual sum of squares (RSS) of each non-parametric regression under series of $\xi \in [1, 40]$ with increment of 0.1, and by least-square selection criteria choosing the minimized RSS (Figure 1.G.3). This criteria gives the best-fitting $\xi = 20.6$, very large and making labor response extrapolation very extreme under low $RMSE$. On the flip side, the peak of RSS exists over the range where the non-parametric arrays are most similar to the main interactive regression at $\xi \in [5, 10]$, discrediting the selection over this range.

With the coefficients estimated for the regressions and the deduced $\beta(w)$ functions again fixing the same reference labor at $25C$, valuation is repeated towards $V(a)$ in Figure 1.G.4. From these plots, the choice of overreacting factor ξ is very significant to the valuation results. When $\xi \leq 5$ the valuation of $V(a)$ is less than 1/10 the magnitude of the main results. When ξ reaches 10, the value inflates to almost 7.5 times the main values and more than 300 times the non-overreaction rational case ($\xi = 1$).

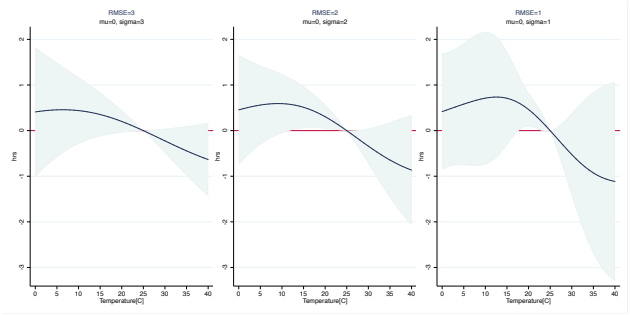
Figure 1.G.2: Restricted Linear Spline Non-Parametric Regression, $\xi = 1, 2, 5, 10$



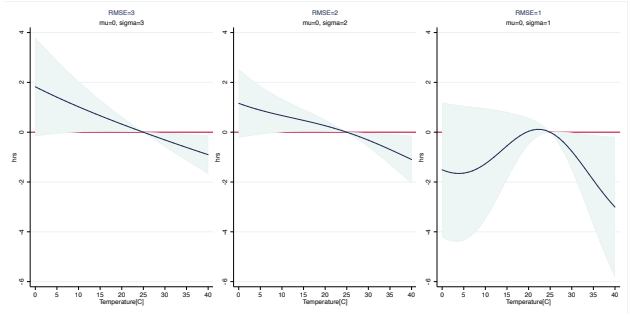
(a) $\xi = 1$



(b) $\xi = 2$



(c) $\xi = 5$



(d) $\xi = 10$

Note: Left to Right, σ decreases $3, 2, 1$, fixing $\mu = 0$.

Figure 1.G.3: Residual Sum of Squares (RSS) of Non-Parametric Regression with Respect to Different Overreaction Factor ξ

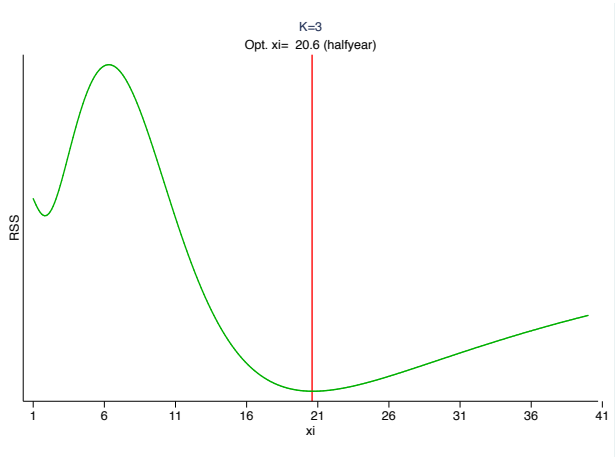
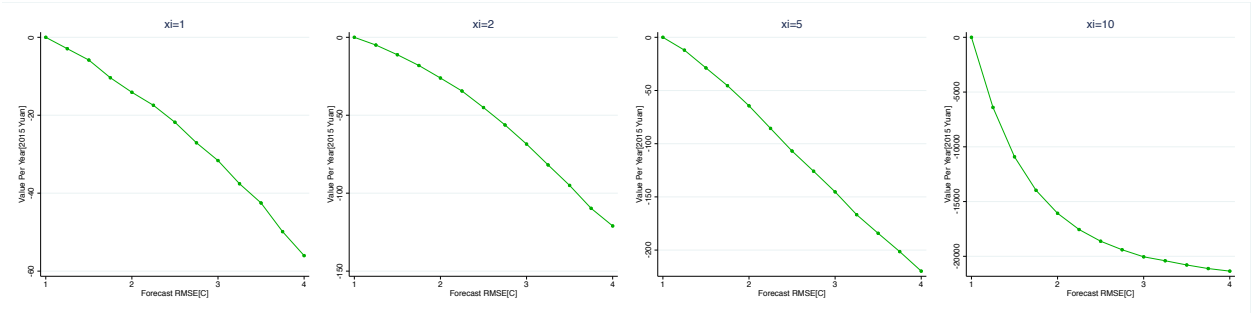


Figure 1.G.4: $V(a)$ Relative to $V(1)$, Per Labor Per Year, Non-Parametrically Estimated with Overreaction Factor $\xi = 1, 2, 5, 10$



1.G.4 Discussions

Overall, this analysis concludes that it is unlikely decision makers form completely rational belief about $w|f, a$, and the valuation following complete rational access of the conditional distribution would result in more than 45 times smaller valuation than my main results. Meanwhile, both estimated labor response and valuation move closer and then greater than the main when people overreact with more than 5 times the true standard deviation of $w|f, a$. As a result, there are some reasonable arguments supporting this overreaction belief model. However, the explanatory power for these regressions with the multiplier from 5 to 10 is lower than the other ranges, not evidence to their picks. Also, there is no sufficient proofs suggesting inflating the belief by more than five times are normal in reality. Therefore, the paper reaches no conclusive evidence on what is people's actual expectation of real temperatures when they receive information about forecasts and forecast accuracy, and this would remain an interesting and important topic for future research.

1.H Other Appendix

1.H.1 National and Labor Sample Comparison

With $N = 11,012$ for my labor sample, there is a usual concern about the external validity problem, namely whether any results from this paper can be extended to conclude cities and dates not covered in the sample. To address this issue, I run tests on whether the labor sample could represents the national population. I compare a selected set of climate and macroeconomic variables using t-tests, between the full year measures for 2011 and 2015

averaged by city and weighted by 2015 city populations, and the summarized labor sample characteristics with sample weights by number of individuals surveyed per city per year. Table 1.H.1 reports the results, where most city characteristics are not statistically significantly different between the labor and national samples as addressed by non-rejection of the t-tests up to 10%. The only difference with significant t-test is that the labor sample cities have more greenland coverage, which among other controls are not very correlated with the labor sector. Therefore, the tests provide partial evidence supporting the labor sample used in this paper being representative of the whole country.

1.H.2 RMSE Breakdown Analysis

By definition, the $RMSE$ metric is related to two components, the mean error (bias) and the standard deviation of the error (uncertainty), on rolling window half a year ($R = 183$):

$$Mean_{it}^{Tmax} = \frac{1}{R}\sum_{s=1}^{R}(Tmax_{it-s}^{real} - Tmax_{it-s}^{forecast})$$
$$STD_{it}^{Tmax} = \frac{1}{R}\sum_{s=1}^{R}[(Tmax_{it-s}^{real} - Tmax_{it-s}^{forecast})^2 - (Mean_{it}^{Tmax})^2]$$

Under law of large numbers, $RMSE \approx \sqrt{Mean^2 + STD^2}$. For comparison purpose, I take the absolute value of the $Mean$ to indicate the size of average forecast error[78]. In my data for all days and all cities with available forecasts 2011-2015, the correlation between $RMSE$ and absolute $Mean$ is 0.945 and for STD is lower at 0.419. The two are not very correlated

[78] 87% of this quantity is negative in the full forecast sample 2011-2015.

Table 1.H.1: Summary Statistics of Labor VS National Sample, Year 2011 and 2015, Population Weighted

Variable	Labor Sample	National Sample	Difference
Daily Forecast Tmax [C]	22.257	21.127	1.131
	(4.348)	(3.612)	[0.450]
Daily Real Tmax [C]	21.004	19.128	1.875
	(4.119)	(3.820)	[0.817]
Daily Precipitation [mm]	2.431	2.694	-0.263
	(1.625)	(1.275)	[-0.722]
City Area [$10^4 km^2$]	1.954	1.756	0.198
	(2.027)	(1.930)	[0.318]
GDP per Capita [10^4 2015 Yuan]	5.568	4.516	1.052
	(2.882)	(2.641)	[1.245]
Road Area [$10^4 km^2$]	0.004	0.003	0.002
	(0.004)	(0.003)	[1.417]
Labor Force/Pop.	0.374	0.256	0.117
	(0.303)	(0.229)	[1.609]
Unemployment Rate [%]	2.592	2.693	-0.101
	(1.404)	(1.296)	[-0.323]
Share of Primary Industry [%]	11.170	12.641	-1.471
	(9.475)	(7.585)	[-0.844]
Share of Secondary Industry [%]	43.889	48.250	-4.360
	(11.616)	(8.376)	[-0.829]
Water Resources ($10^8 m^3$)	90.799	75.896	14.903
	(133.222)	(95.132)	[0.380]
Greenland Area [$10^4 km^2$]	0.027	0.010	0.017
	(0.038)	(0.020)	[2.281]**
Elevation [m]	286.703	418.824	-132.122
	(314.202)	(507.991)	[-1.525]
No. Weather Stations	2.304	2.056	0.248
	(2.458)	(1.861)	[0.337]
Area of Lakes [$10^4 km^2$]	0.035	0.030	0.005
	(0.057)	(0.068)	[0.516]
Lenghth of Rivers [Deg]	47.993	39.905	8.088
	(40.500)	(37.099)	[0.647]

Note: Labor Sample has 52 cities and only labor reporting weeks, national sample covers all 342 over all days of 2011 and 2015; All variables are averaged by city; GDP per capita adjusted to 2015 by World Bank GDP deflater; Labor sample are weighted by number of individuals in each city; For national sample, city averages are weighted by 2015 population; Standard deviations are in parenthesis; Difference has t-test performed and t-statistics included in square bracket; For t-test results, $^{*}p < 0.1$, $^{**}p < 0.5$, $^{***}p < 0.01$.

between themselves, with a correlation coefficient of 0.151. But STD has twice the average size as the absolute $Mean$, and in terms of squares, STD contributes on average 68.4% of the $RMSE$ square, while $Mean$ only 32.0%. In summary, this mean that $Mean$ likely has more similar spatial and temporal trend as $RMSE$, but STD contributes greater to the value of $RMSE$.

With summary plots Figure 1.H.1, absolute $Mean$ and STD features quite differently overall. From Panels (a) and (b), absolute $Mean$ has smaller average across both samples and greater distribution variation than STD. Both absolute $Mean$ and STD components have positive skewness for either sample choice. In Panels (c) and (d), both features large spatial variations, but their correlation is not perfect and neither has great proximity to the map of $RMSE$. For example, south-west regions have relatively small absolute $Mean$ but greater STD. Also by construction of forecast adjustment for non-capital cities, absolute $Mean$ varies much more across province borders than within same province, which does not exist for STD.

More difference between the two occurs for temporal variations in Panels (e) and (f). Absolute $Mean$ increases by 2.3% while STD decreases by 5.1% from 2011 to 2015, implying that the forecast accuracy improvement is largely contributed by lowered uncertainties. However, the magnitudes of both changes are still small, only $0.044C$ for absolute $Mean$ and $0.124C$ for STD. Like $RMSE$, absolute $Mean$ maximizes at fall and minimizes at spring, while STD has very different monthly trend with maximum around summer (therefore first half of the year has smaller forecasts precision). For yearly trend, absolute $Mean$ sees more

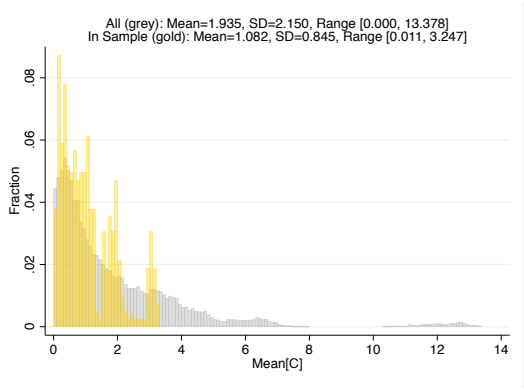
decrease among spring months but offset by the increase over fall months, while STD uniformly decreases for all months from 2011 to 2015[79].

So overall, there are improvement in forecasts precision over the 5 years I study. However, temporal variations still contribute little comparing with spatial variations for both components (for absolute $Mean$, spatial contributions 91.5% for all sample, 62.2% for labor sample; for STD, spatial contributions 80.0% for all sample, 72.7% for labor sample), making any analysis using them instead of $RMSE$ still focusing on spatial variations in forecast accuracy. The choice of using $RMSE$ instead of these two separate components are based mainly on the hypothesis that people are concerned about both the bias and the precision of weather forecasts. As a verification, the explanatory power of the baseline interactive regression is greater than using the two breakdown components to represent forecast accuracy instead (see robustness checks Appendix Figure 1.D.16).

[79]Though there are fluctuations in earlier years (for example, STD increases for some months 2011 to 2012).

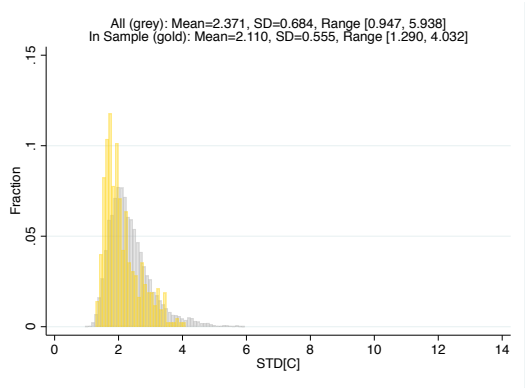
Figure 1.H.1: Summary Plots for the Spatial and Temporal Variation of Forecast RMSE Breakdown

(a) Daily Maximum Temperature Absolute Forecast Error Mean on Rolling Window of Half a Year (183 Days)



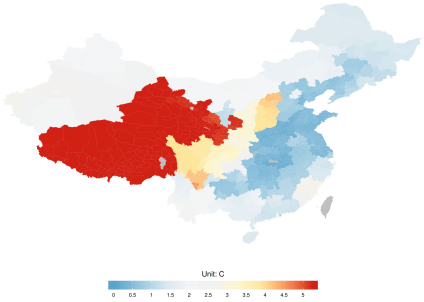
Note: Gold is only the labor sample coverage, Grey is all 342 cities of all days 2011 and 2015 with non-missing forecasts (685 days).

(b) Daily Maximum Temperature Forecast Error STD on Rolling Window of Half a Year (183 Days)



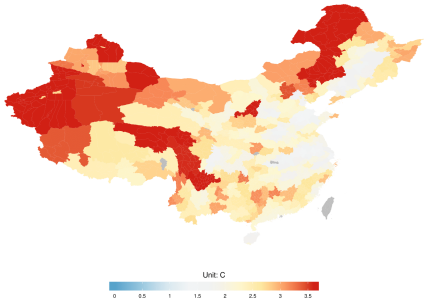
Note: Gold is only the labor sample coverage, Grey is all 342 cities of all days 2011 and 2015 with non-missing forecasts (685 days).

(c) Half-Year Rolling Absolute Forecast Error Mean Average Across Time By Cities



Note: All 342 cities, over 685 days with non-missing forecasts in 2011 and 2015.

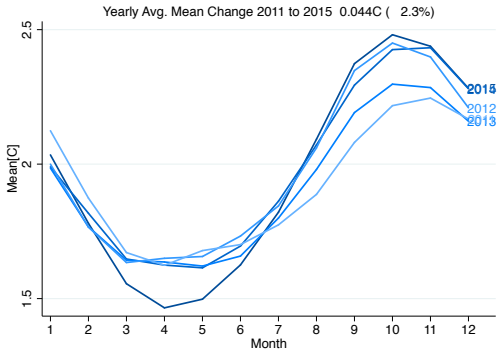
(d) Half-Year Rolling Forecast Error STD Average Across Time By Cities



Note: All 342 cities, over 685 days with non-missing forecasts in 2011 and 2015.

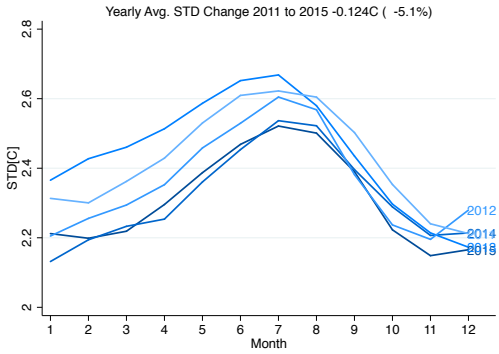
Figure 1.H.1, continued

(e) Half-Year Rolling Absolute Forecast Error Mean Average Across Cities By Months



Note: Monthly trend plots by years 2011-2015, sample includes all 342 cities over all days with non-missing forecasts.

(f) Half-Year Rolling Forecast Error STD Average Across Cities By Months



Note: Monthly trend plots by years 2011-2015, sample includes all 342 cities over all days with non-missing forecasts.

Note: Left, Absolute Mean; Right, Standard Deviation; Both on Rolling Window R=183 Days.

Chapter 2

The Value of Accurate Weather Forecasts: Social Sentiment Responses Reflected in Social Media in China

Abstract

This paper combines real weather data with the city-daily weather forecasts broadcast by the government, as well as the sentiment index expressed by posts on the popular social media Weibo in China, and through an interactive regression design analyzes the differential sentiment responses to temperatures under different sizes and signs of daily temperature forecast errors. My main results have suggested that more accurate temperature forecasts lead to smaller shifts towards unhappiness caused by the cold temperatures. The same effects does not play out under hot weathers, unless in cities with higher income, greater long run temperature forecast accuracy, or during holidays. My study also suggests that additional negative sentiment shocks are likely related to cold or heat alarms issued according to the national forecasts, resulting in that positive forecast errors have greater marginal effects on sentiment than negative errors during the cold temperatures. Overall, these results meet the intuition that advanced forecast technology provides more accurate daily temperature forecasts, and adds to great social benefits in China in terms of improving people's subjective well-beings as expressed by social media sentiments. In the current time under climate change, when extremal weather events are expected with greater frequency into the future, my work would help to provide an insight of the value of developing a modern weather forecasting system that can benefit billions of people in the long run.

2.1 Introduction

It has been known in economics, that subjective well-beings, sometimes expressed as people's happiness of life, are highly correlated but not equivalent to measurable economic

standards such as income and consumption levels (Luigino, 2004). Other less quantified factors are also instrumental to the happiness index, such as health and general living standard (Glatzer, Camfield, Moller and Rojas, 2015; Carleton etal., 2020). In modern days with the wide usage of social media all across the globe, the traditional method of surveying to measure the subjective well-beings have been replaced by text analysis using machine learning natural language processing (NLP) (Kahneman and Krueger, 2006; Dodds etal., 2011). Through the fast growing literatures using social media expressed sentiment index representing people's real-time subjective well-being, there have been many studies on contributors to this happiness metric. For example, pollution, extreme temperatures, and natural disasters are all found to be important factors for people's sentiment on social media (Zheng etal., 2019; Wang, Obradovich and Zheng, 2020; Kryvasheyeu etal., 2016). However, there has not been much explorations on one factor that can potentially impact the society happiness level in the long run, namely the technology development. In fact, technology development has positively impact a range of known factors contributing to people's utilities, including economic growth, quality of living, health improvement, and risk reduction encountering natural disasters.

In this paper, I want to focus on the modern technology of weather forecasting, one of the most common technology available to most population across the globe as a free public good. Nowadays, sophisticated weather forecasting systems managed by governments and research institutions are combinations of numerical modeling and professional judgment, guided by most advanced development in the fields of meteorology, engineering, statistics, and computer science. Over the past decades, large amount of financial supports have been

invested in predicting more accurate weather forecasts, from generating new numerical forecasting models to launching more meteorological satellites. Scientists and engineers have been working on providing the best forecasting systems to ensure the timely and reliable predictions of future weathers, from days ahead to weeks ahead and even years to come (Edwards, 2011; Coen, 2021). From general public to governments, more accurate forecasts help all these decision makers broadly with preparation to avoid potential damages caused by future unpleasant weather events.

There have not been a large literature on how valuable is "good" weather forecasting to the society. Previous studies have mostly focus on the medium-run precipitation forecasts provided to smaller groups of decision makers in specific industries. For many of those, historical forecast data is not available, so authors rerun numerical forecasting models or use other first stage predictions. Most of them have found negative impacts of the failure of forecasts on agriculture and fishery productions as well as related corporate revenues (Allen, Graff Zivin and Shrader, 2016; Shrader, 2020; Rosenzweig and Udry, 2019; Downey, Lind and Shrader, 2021). For a further step, quantifying the value of weather forecasts has been non-trivial. Conventional methodologies applied have usually included contingent valuation, which surveyed the benefit of weather forecast systems from several billions to more than 30 billion USD depending on the country and time of studies (Yuan, Sun and Wang, 2016; Lazo, Morss and Demuth, 2009). Otherwise economic approaches have been applied to evaluate the impacts of accurate forecasts on transportation, economic damages and production yields, giving estimates for the value of forecasts again from several billions to more than 80 billions (Nurmi etal., 2012; Martinez, 2020; Fox, Turner and Gillespie, 1999). Though their

settings and methods differ, all of these studies have found a large benefit to cost ratio for the weather forecasting systems around the world.

This paper instead focuses on the less studied short-term temperature forecasts, which are easily accessible by the larger population. With the choice of setting in China, where the nationwide weather forecast system has been managed by one state agency, China Meteorological Administration (CMA), this project studies the 24-hour daily forecasts provided by an almost uniform source. For the outcome variable I choose to look at the sentiment index expressed on one of the largest social media platform in China, Sina Weibo, since the day-to-day sentiment response is likely sensitive to daily weathers, and whether those weathers are forecast accurately. Combining two unique datasets in the year of 2014, I use the climate economic approach to directly evaluate the impacts of accurate temperature forecasts on social well-beings in China. In summary, my results provide evidence that more accurate temperature forecasts help to improve people's sentiment in real time. To be specific, greater daily temperature forecast errors have contributed to greater unhappiness under cold weathers, especially when forecasts give exaggerated cold warnings. Meanwhile, the negative sentiment shock in response to heat has not been affected significantly by these instantaneous forecast errors. The discrepancy between cold and hot sides may be contributed by the higher cost of adaptation during heat.

This paper proceeds as the following. Section 2 describes datasets, data treatments and summary statistics. Section 3 outlines the empirical design and then presents the main results. Section 4 conducts extension explorations and robustness checks. Section 5 concludes.

2.2 Data and Summary Statistics

2.2.1 Sentiment Index

I received the unique sentiment index dataset for Chinese social media from MIT Substainable Urbanization Lab (SUL), which has been applied in their papers Zheng etal. (2019) and Wang, Obradovich and Zheng (2020). As described in those literatures, they scraped 210 million microblog tweets with geotags (i.e., the authors microblogs choose to identify their city) posted on the Chinese largest microblog platform Sina Weibo (similar to Twitter). Then, the "Tencent" natural language processing (NLP) platform is used to code a machine-trained sentiment analysis algorithm from computational linguistics. This algorithm is applied to measure the sentiment for each Weibo post, with which post is given an index ranges from 0 to 100, where 0 means a strongly negative ("Unhappy") and 100 a strongly positive ("Happy") mood. Next, the overall happiness index for a city is constructed by calculating the mean and median sentiment value of the sentiment indices across all the geotagged posts generated on a given day. Eventually, the SUL sentiment index data covers 144 Chinese cities from all 31 mainland provinces, over the period of 275 days from March to November 2014 (all seasons except for winter).[1] The city mean sentiment index in this dataset spans a range from 35.7 to 79.9 with an average of 55.2 and standard deviation 2.4. Day-to-day shift in city mean sentiment index can be pretty large, ranging from −24.2 to 21.5, though the average shift is fairly close to zero.

[1]For the specific purpose related to their pollution studies, the lab uses a dictionary of pollution terms to exclude the microblogs discussing air quality. The proportion being excluded is fairly low, about 0.047% of all the posts.

2.2.2 Weather Forecast Data

In China, the national weather forecasting system has been managed by the state department of China Meteorological Administration (CMA), which is in charge of both producing and broadcasting the national forecasts. Their forecasting job include several steps. First, multiple numerical models take inputs including historical and current weathers from observatory stations and weather satellite images, considering parameters like local geographic factors (for example, elevations), and output weather forecasts (including temperatures and precipitations) for the near future (usually 2-3 days)[2]. Next, professional weather forecasters trained and employed at local weather stations evaluate the raw predictions from different models, and summarize the "average" forecasts with their personal experience and judgment taken into considerations. Lastly, CMA holds the daily conference with local forecasters across the country, collects their final local forecasts, verifies their consistency at a national level, and distributes the final forecasts to the public via TV, radio, papers and Internet[3].

The almost uniform source of weather forecasts from CMA in China has ensured a scenario that most of the Chinese population receives the same forecast information [4]. According to Yuan, Sun and Wang (2016), over 40% population in China receive weather forecast

[2]I consult the CMA, they did not specify which models they used, but confirm there are more than one models involved and their list, and the list has been updated over the years. When old models are dropped some will be posted on their website, but new models are being added and modified all the time.

[3]News source http://www.cma.gov.cn/2011xzt/2013zhuant/20130524/, https://zhuanlan.zhihu.com/p/21598589.

[4]Many third-party websites and mobile apps also adapt weather forecasts, but majority of them quoting the information from CMA. One common app not quoting CMA is the Apple weather app. However, though people may rely on apps more for real time temperatures, in general they still tend to rely on the authority source for weather forecast information.

information from TV in 2006, the dominant source among all age groups. And among all forecasts, the 24-hour weather forecast is one of the most popular and most emphasized products. This information includes temperature forecasts of a range and a weather category forecast (e.g., sunny, shady, small rains, fog) for cities across the country for the next day to come. It is aired in one of the highest viewership TV programs of the country, the *Weather Forecast*. This program is aired every day on CCTV Channel 1 (China Central Television - Main Channel) and Channel 13 (News Channel). Over the past decades, the program has expanded to three times a day, in the morning, at noon, and in the evening. The evening *Weather forecast* receives significantly higher viewership comparing with the other two, especially among elder generations, because it is aired almost immediately after the daily national news ending at 7:30pm[5].

I extracts these next-day weather forecasts from this popular evening *Weather Forecast* program as information perceived by the majority of audience. To do so, I download the full videos published by the official website of CCTV.com, which is real-time recording of the program broadcast on TV and made available on the Internet by the end of each day. Due to a copyright constraint, videos on the website only cover back to the year of 2010. I batch download all videos available (some days especially in earlier years are missing) using the video consolidating-downloading site FLVCD. The downloaded videos are then transformed to FLAC audio files with fitting wavelength numbers, and fed into the Google Cloud speech-to-text API. The API reads the audio files and transcribe speeches to Chinese texts in TXT

[5]The intermission between the news and weather forecasts is around 1 minute, filled by only a couple commercials. This intermission period has been reported as one of the most expensive TV advertising slots because of high viewership.

scripts. Finally, I clean up the scripts with STATA to collect observations of forecasts identified by city and date, and those are finally compiled into my Chinese weather forecast dataset.

This weather forecast dataset contains the temperature range forecasts in the form of $Tmin$ and $Tmax$, as well as the categorized weather forecasts in the next 24 hours for all 34 provincial capital cities in China. I take daily average temperatures as the mid point of the temperature range, $Tavg = \frac{1}{2}(Tmin + Tmax)$. To expand to all Chinese cities, I approximate temperature forecasts for non-capital cities by adjusting a provincial capital forecast with a difference between the monthly(m) average real temperature (source ERA-Interim, see next subsection) of non-capital i and its capital city p in 2010:

$$T_{itmp}^{forecast} = T_{tmp}^{forecast} + \frac{1}{\sum_{year=2010,month=m} \mathbf{1}} \sum_{year=2010,month=m} (T_{itmp}^{real} - T_{tmp}^{real})$$

And the categorized forecasts for those non-capital cities are approximated by the same categorized forecasts as their provincial capitals. This would be relatively rough approximations comparing with the temperature forecasts, however.

In the end, this weather forecast dataset then cover all cities in the sentiment dataset. That includes 31 provincial capitals which receive direct forecasts, and 113 non-capital cities with approximated forecasts. I take the time range to be the year of 2014, which overlaps with the sentiment dataset.

2.2.3 Real Weather Data

The "real" climate data available for years 2009 to 2016 is sourced at ERA-Interim (ERAI) reanalysis data product from ECMWF (European Centre for Medium-Range Weather Forecast). This dataset covers daily temperatures $Tmin$, $Tmax$ (and $Tavg$ again taken as the average), and precipitations in mm. Strictly speaking, this data product is neither the real-time recordings nor the raw historical weather station readings, but instead extrapolations using mathematical models on existing station recordings taken with time intervals. Their models approximate weathers to a high frequency of 3 hourly, 0.25×0.25 grid level. Normally, climate scientists judge this data product as efficiently close to the real historical weathers, especially for temperatures, though there are more uncertainties around the extrapolation of precipitations. Eventually, I source both this ERAI data product and the generating codes from EPIC (Energy Policy Institutde at the University of Chicago) Climate Impact Lab (CIL), to aggregate the ERAI real weather data to city-daily levels by population weights, in order to match with the frequency of the forecast dataset.

2.2.4 Control Variables

Various environmental and socio-economic variables are used to separate subsamples analysis or testing of exclusion restriction. Specifically, main city and economic indicators are obtained from China City Statistical Yearbook, including GDP per capita, population, road area, green land coverage, labor force breakdown, industrial pollutant emissions, water and electricity supply, coal and petroleum consumption, city areas and water resources. City

boundaries, lake and river distributions are aggregated using the GDB shapefiles applied from National Catalogue Service For Geographic Information, published year 2017, released by State Bureau of Surveying and Mapping. City elevation data is obtained from Appendix table of the Load Code for Design of Building Structure (GB 50009-2012). The location of weather stations in China is obtained from records of NOAA (National Oceanic and Atmospheric Administration) Integrated Surface Database (ISD). I also obtain air pollution data scraped from Ministry of Ecology and Environment of the People's Republic of China (daily, 2013-2018) or the Anthropogenic Aerosol Optical Depth (AOD) (monthly, 2005-2015) downloaded from NASA website.

2.2.5 Rationalization of Temperature Forecasts

In real life, it would be hard to argue how general public form their beliefs on tomorrow's weather after receiving the temperature forecasts today. For example, if the forecast has been consistently off by $1C$, will residents notice the constant error and induce an accurate forecast by adding it back, or will they take the raw forecasts as given. This assumption not only affects whether people respond to raw or adjusted forecast information, but also affects their perceived forecast accuracy.

An alternative to raw forecasts would be the "rationalized" forecasts. By definition, a rational forecast is a forecast with symmetrical forecast error centering around 0. Under symmetric loss function, it follows that positive and negative forecast errors can be weighted equally. To test whether the weather forecast data I have is rational, I follow the theory in

Mincer and Zarnowitz (1969) and run the following simple OLS with all 12 months of year 2014 (including the winter months where the sentiment sample is not spanning) with robust standard error:

$$T_{it}^{real} = \alpha_{iy} + \beta_{iy} T_{it}^{forecast} + \epsilon_{it}$$

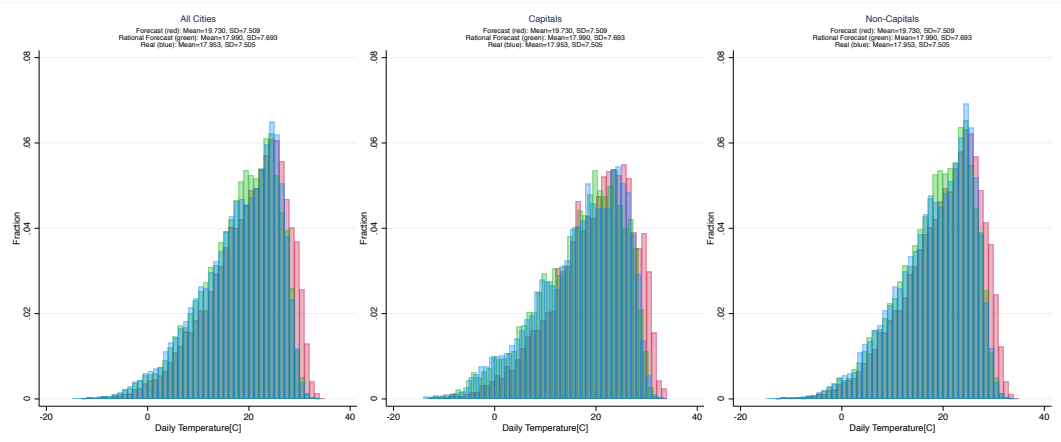
Rational forecasts require the null $\alpha_{iy} = 0$ and $\beta_{iy} = 1$ to be not rejected. Running the F-test for all 144 sentiment index cities individually, over 99% of the p-values for all $Tmin$, $Tavg$ and $Tmax$ falls below 5% (over 98% are rejected at 1% level). Therefore, vast majority of the null are rejected, so that the raw forecasts are not rational.

As a result, I would consider this alternative hypothesis that individual adopt rationalized forecast instead of the raw, where the "rationalization" of raw forecasts is conducted with individuals adjust the raw forecasts linearly with the OLS estimates achieved above:

$$\hat{T}_{it}^{forecast} = \hat{\alpha}_{iy} + \hat{\beta}_{iy} T_{it}^{forecast}$$

In Figure 2.1 it illustrates the daily $Tavg$ distribution across all cities, only capital cities and only non-capital cities. From the histograms we can see similar shapes of distributions for real, non-rationalized and rationalized forecast temperatures regardless of the sample, verifying that the forecast approximation process I proposed for non-capital cities has been reasonable. In general, the non-rationalized forecasts (red) are higher than the real temperatures, but the rationalization process seems to correct it such that the rationalized forecast (green) overlaps with the real temperature.

Figure 2.1: Daily Average Temperature $Tavg$: Blue=Real, Red=Non-Rationalized Forecast, Green=Rationalized Forecast; Covering the Sentiment Sample March-November 2014; Left to right samples: all 144 cities, only 31 capital cities, only 113 non-capital cities.



To encounter both assumptions regarding how the public take forecast information into beliefs, for the later sections, results will be run with both raw "non-rationalized" forecasts as well as these "rationalized" forecasts.

2.2.6 Forecast Error Definition

According to the model proposed in Hsiang (2016), for outcome variables sensitive to weathers like the sentiment, there exists responses to both real weathers and weather expectations. In the case of sentiment, which in nature has a fast-changing characteristic (e.g., sentiment can alter from minute to minute), response to real weathers is likely dominant. However, such sentiment response to real weathers is also likely related to how well bad weather events have been forecast and prepared for based on the accuracy of those forecasts. For example, a reasonable hypothesis is that there will be negative sentiment shocks corre-

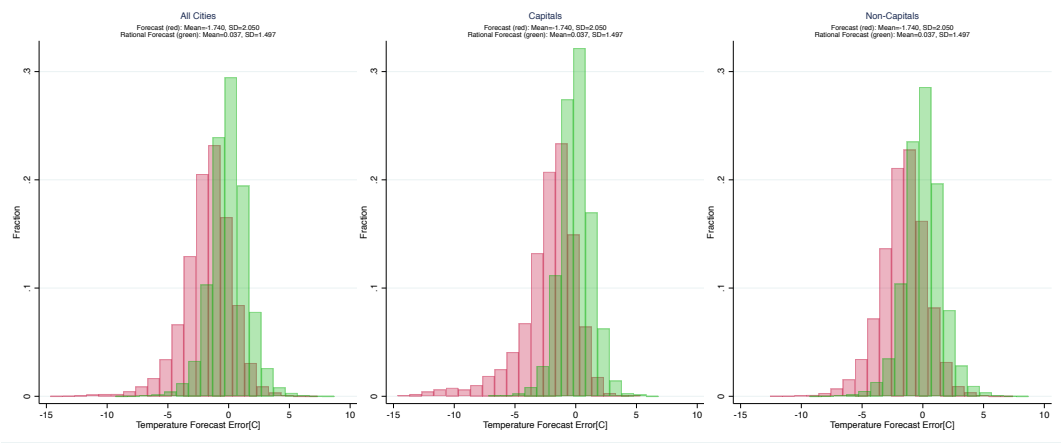
sponding to bad weathers, but the magnitudes of those negative shocks would be lowered by providing better quality weather forecasts.

In my setting, the Chinese national forecast system has been established since the 1980s, and weather forecasts nowadays have become quite accurate thanks to technology development. However, no forecasts can consider all the real-time factors perfectly, giving the rise of instantaneous forecast error:

$$T_{it}^{real} - T_{it}^{forecast}$$

Here T is a daily temperature measure, $Tmin$, $Tmax$ or $Tavg$. This error metric represents how accurate the forecast is for city i at day t. By definition, it can be either positive or negative. A positive forecast error represents an underestimate of the temperature, and a negative forecast error represents an overestimate. In Figure 2.2, we can see mainly negative errors with non-rationalized forecast (red), but symmetrical error distribution of the rationalized forecasts (green) by design. Notably, there exists extreme errors with great magnitudes $> 10C$ on both positive and negative ends for the non-rationalized forecasts, for which I have checked to be over the western part of the country where mountainous topography likely makes weather forecasting more difficult and inaccurate.

Figure 2.2: Forecast Error for Daily Average Temperature $Tavg^{real} - Tavg^{forecast}$



Note: Blue=Real, Red=Non-Rationalized Forecast, Green=Rationalized Forecast; Covering the Sentiment Sample March-November 2014; Left to right samples: all 144 cities, only 31 capital cities, only 113 non-capital cities.

2.2.7 Spatial and Temporal Variations of Forecast Accuracy

Based on the process of Chinese weather forecasting, theoretically, forecast errors are relevant to a series of observable and unobservable factors. Errors can due to missing and mis-measured observational inputs for the numerical models (for example, errors in current weather recording, inability to capture certain airmass movements with restriction to satellite access), systematic prediction errors of modeling (for example, old models do not capture the new earth dynamics under climate change, high elevation weathers are usually difficult to predict due to fast and complex airmass movements), and human errors when forecasters summarize the model outputs (for example, sincere mistakes or intentional tampering of data). As a result, this metric is expected to vary both in time and in space. Spatially, geographical parameters (e.g., nearby a water body or inland, high versus low elevation) and the ability of local weather forecasters varies. Temporally, numerical models and professionals may be better at predicting weathers under specific weather conditions during some seasons

but not others.

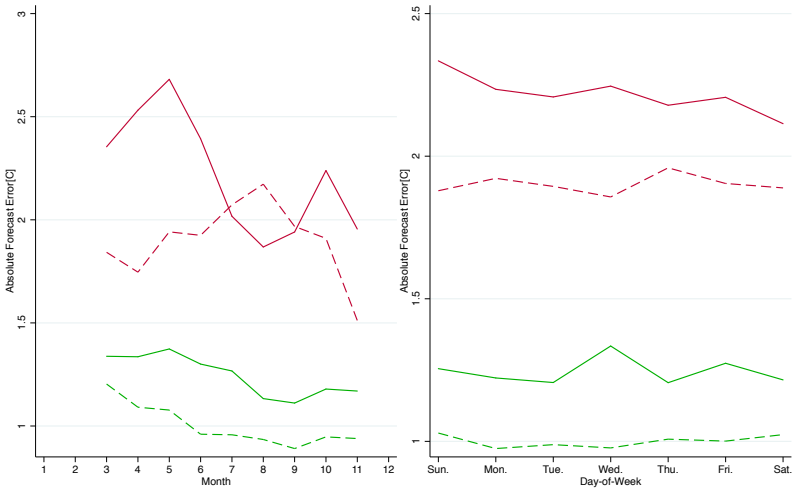
In Figure 2.3, I show some instances of the temporal variation of the magnitudes of this forecast error with $Tavg$. As it shows, there seems to be decreasing monthly trends of average $|Tavg^{real} - Tavg^{forecast}|$ (left) approaching end of year 2014, for both northern (solid) and southern (dash) part of China, for both non-rationalized (red) and rationalized (green) forecasts. There is less of a visible time trend over the day-of-week average (right) however, implying that the forecasting system is carried out identically throughout the week with small likelihood of human errors being dependent on weekdays or weekends.

On the other hand, spatial variations within and across province boundaries can be seen from the yearly average absolute forecast error for $Tavg$ in Figure 2.4. Especially for raw forecast (left), large average errors seem to be persistent in the northern and western part of the country, which are inland and mountainous. This observation overlaps with the extremes of daily forecast errors noted in previously. But after rationalization, such outliers disappear and average forecast errors are more similar across different regions (right).

2.2.8 Exclusion Restriction Assumption

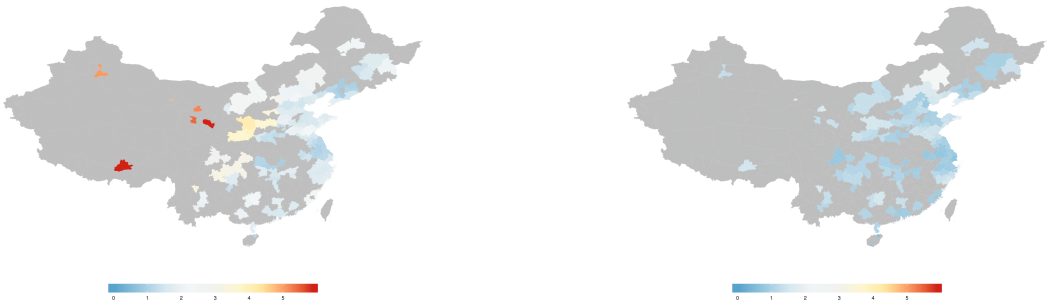
The exclusion restriction hypothesis, namely the impacts of temperature forecast error $T^{real} - T^{forecast}$ on the sentiment responses to temperatures is not caused by other related factors, is important for my argument that forecast and forecast accuracy plays a significant role in sentiment shocks related to weathers. There is no sufficient testing for this

Figure 2.3: Daily Average Temperature Absolute Forecast Error $|Tavg^{real} - Tavg^{forecast}|$



Note: Red=Non-Rationalized Forecast, Green=Rationalized Forecast; Left: Average by Months, Right: Average by Day-of-Week; Solid(North) vs Dash(South).

Figure 2.4: Maps display the average absolute daily forecast error $|Tavg^{real} - Tavg^{forecast}|$ **for each city across the sample period of March-November, 2014.**

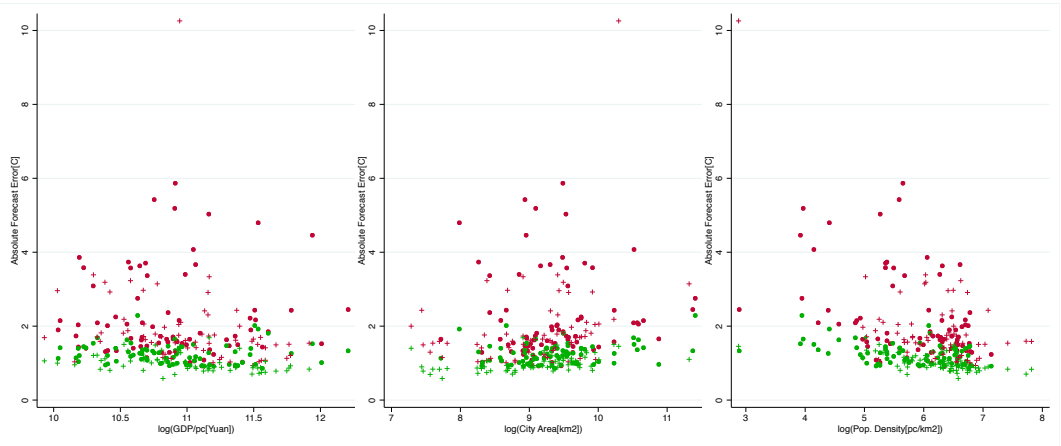


(a) Non-Rationalized Forecast Error

(b) Rationalized Forecast Error

hypothesis though. Instead in this part, I would argue that the forecast error metric is not fully generated by a series of observed exogenous factors. These factors can be those inputs of numeric modeling like geographic variables, historical and current climate conditions, or factors related to human errors of local forecasters such as city level socia-economic variables[6].

Figure 2.5: Daily Average Temperature Absolute Forecast Error $|Tavg^{real} - Tavg^{forecast}|$



Note: Red=Non-Rationalized Forecast, Green=Rationalized Forecast, Averaged by Cities; Left: Scatter with Income, Middle: Scatter with City Areas, Right: Scatter with Population Density; Dots(North) vs Cross(South).

In Figure 2.5 I present some simple scattering analysis of three factors that may have been related to forecast accuracy and sentiment shocks simultaneously. For all these factors, I see a scatter of the average absolute forecast error across northern and southern cities with different GDP per capita, city area and population density. There are some correlations that matches with intuitions, for example positive with city area (larger city make forecasting harder to summarize over the greater area), and negative with population density (lower population density cities are likely to be in mountainous areas where numerical modeling

[6]Human errors may be results of intentional manipulations, but it still would be less likely with the heavy dependence of exogenous numeric modeling outputs and the validation check administrated by CMA.

finds it harder to forecast), but they seem not able to explain all the scattering.

With these arguments, I further run the multivariate OLS with city level factors to predict the forecast error metric:

$$T_{it}^{real} - T_{it}^{forecast} = \alpha + \beta' \mathbf{X}_{it} + \gamma_i + \delta_t + \epsilon_{it}$$

Where a selected set of controls $\mathbf{X}$ is chosen including the city area, amount of water resources, area of city greenland, GDP per capita, population, total road area, ratio of labor over population, unemployment rate, share of primary, secondary and tertiary industry, industrial dust emission, city elevation, number of weather stations, area of water bodies, total length of rivers, mean and standard deviation of real daily temperatures, mean daily precipitation over the period 2002-2011, and average monthly AOD over period 2005-2015. City and date fixed effects are also included.

Conducting 10-fold cross-validation for the regression, the out-of-sample R-squared (pseudo R-squared) for this predicting regression is about 30%-40% for non-rationalized forecasts, and only 10%-13% for rationalized forecasts. Overall, this shows that the forecast error metric is not predicted by these possible relevant control factors, with predicting power lying below 50%. Thus as addressed by exclusion restriction, I may expect the forecast error to contribute as the covariate for the sentiment response to daily temperatures.

2.3 Empirical Design and Main Results

2.3.1 Main Interactive Regression

Now with city-daily temperature forecast errors as the perceived weather forecasts accuracy metric, my main goal is to estimate the differential sentiment responses to temperatures under different forecast accuracy. The empirical design I adopt follows from setting in Carleton etal. (2020), it is an interactive regression treating the absolute instantaneous forecast error as a linear covariate, and allowing the interactive terms to differ for negative and positive forecast errors separately:

$$\begin{aligned} Sentiment_{it} = \alpha + f_0(T_{it}^{real}) \\ + f_-(T_{it}^{real}) \times \mathbf{1}_{T_{it}^{real} - T_{it}^{forecast} < 0} \times |T_{it}^{real} - T_{it}^{forecast}| \\ + f_+(T_{it}^{real}) \times \mathbf{1}_{T_{it}^{real} - T_{it}^{forecast} \geq 0} \times |T_{it}^{real} - T_{it}^{forecast}| \\ + \gamma' \mathbf{X}_{it} + \epsilon_{it} \end{aligned} \tag{2.1}$$

Note index i is the city where forecast and sentiment reflects, t is the day. For here, $T^{forecast}$ is the 24-hour day-ahead forecast broadcast in day $t-1$ reporting temperature forecast of the next day t. The outcome variable is taken to be the mean sentiment at i in day t, and regression is weighted by population of each city in 2014. In this design, the direct temperature response f_0 is common for any forecast errors. I allow for different interactive responses with absolute forecast errors $|T^{real} - T^{forecast}|$ when the error is positive or negative, f_+ and f_-. The non-linear response functions f, f_+, f_- all take the same forms,

in the main analysis they would be non-parametric $5C$ temperature bins (regression with series of bin indicators B_k, each indicating the number of days where the temperature falls in the bin range $T_{it} \in B_k$) with a reference bin dropped. Controls $\mathbf{X}_{it}$ include real precipitation and square, day and city fixed effects (FE). Standard errors are clustered at city level.

This interactive regression would estimate the non-linear temperature response for sentiment:

$$\widehat{Sentiment}(T, Error)$$

Where $T = T^{real}$ is real temperatures, $Error = T^{real} - T^{forecsat}$ is the instantaneous forecast error. In the analysis, these results will be illustrated with panels of $\widehat{Sentiment} - \widehat{Sentiment}^{ref}$ against T under given $Error$, where $\widehat{Sentiment}^{ref}$ is taken under the reference temperature T^{ref} supposed to be "comfortable" for human. To interpret, this would be the trend of estimated change in sentiment at real temperature relative to the reference temperature given a forecast error. In my bin regression, T^{ref} is the bin omitted and $\widehat{Sentiment}^{ref} = 0$. The plots will be taken for a series of forecast errors $T^{real} - T^{forecast}$ from negative to zero then to positive. Also, plots for the marginal effects (ME) of covariate $|T^{real} - T^{forecast}|$ will be presented for both negative and positive interactions, which are $\hat{f}_+(T) - \hat{f}_+(T^{ref})$ and $\hat{f}_-(T) - \hat{f}_-(T^{ref})$ against T. These marginal effect (ME) plots illustrate the estimated change in sentiment per $1C$ increase of error $|T^{real} - T^{forecast}|$ under forecast T relative to T^{ref}.

2.3.2 Global Regression Check

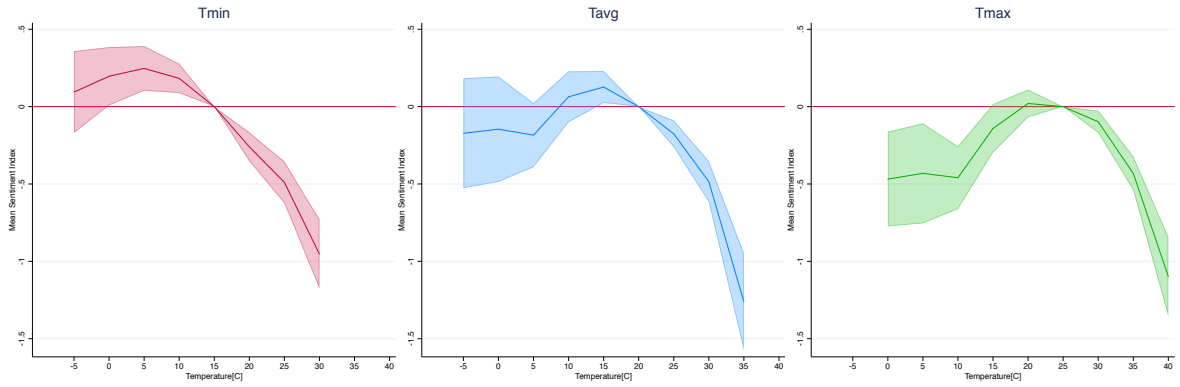
Before going into the main results, I want to check the global response losing the covariate part with forecast errors:

$$Sentiment_{it} = \alpha + f(T_{it}^{real}) + \gamma' \mathbf{X}_{it} + \epsilon_{it} \tag{2.2}$$

This is the empirical design applied in the previous literature using this sentiment index data, Wang, Obradovich and Zheng (2020). The results are presented in Figure 2.6. Here we see inverse U-shape sentiment response to temperatures for all three panels, though the cold end response is only negative and statistically significant for $Tmax$. Comparing with Wang, Obradovich and Zheng (2020) which uses $Tmax$, the results are pretty similar, showing negative sentiment shocks during temperature extremes and the shock is greater for hot instead of to cold temperatures. The size of my estimates are smaller than the paper's, but of same order of magnitudes. In general the discrepancy may be because their paper uses the raw post-daily sentiment data while I apply city-daily average sentiment with population weighting. Also the source of our real temperatures differ.

Overall, these results show that my regression is similar to the previous paper using the same data. This verifies the choice of the functional form as well as the hypothesis of existing sentiment negative responses to extreme hot or cold temperatures. Another important implication is that the negative sentiment shock at the cold side is most prominent for $Tmax$ comparing with $Tmin$ and $Tavg$ (the hot side responses is large and significant for all three, but the magnitude is greatest for $Tmin$ instead). To explain, it can be that

Figure 2.6: Global Regression of Sentiment Responses to Tmin, Tavg, Tmax



Note: Left to Right: T being $Tmin, Tavg, Tmax$ for real temperature; Reference Bins: $[10C, 15C), [15C, 20C), [20C, 25C)$ respectively; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

people in cold weathers form their sentiment around the best temperature during the day, or simply because $Tmax$ is recorded during the midday when social media is most active, while $Tmin$ is usually at midnight when people are mostly offline. However, taking $Tmax$ only into account may still ignore the impact of daily minimum temperatures on sentiments, especially under cold. Therefore, I will mainly take sentiment response to $T = Tavg$, which is the average of $Tmin$ and $Tmax$, in my following analysis.

2.3.3 Interactive Regression with Tavg

The main interactive regression results showing sentiment response to daily average temperatures are shown in Figure 2.7, for both non-rationalized (gold) and rationalized (teal) forecast errors. From the last two columns of marginal effects of this plot, a clear observation is that there is negative significant marginal effect of absolute forecast errors when temperature is low (approximately $Tavg < 5C$), regardless of errors being positive or neg-

ative. However, the ME at the hot end is close to zero (except for negative rationalized forecast errors) and statistically insignificant. The same has been reflected in the first seven columns, whenever the size of instantaneous forecast error decreases, no matter from negative to zero or positive to zero, the negative sentiment response under cold temperatures has been decreasing to statistically insignificant and close to zero. Meanwhile on the hot end, the negative sentiment shock in response to heat has always been sharp and significant regardless of the size or sign of instantaneous forecast errors of $Tavg$.

In interpretation, people on social media are more unhappy about cold weathers when the cold temperature forecasts is not that accurate. This is in line with the intuition that better forecasts can better prepare people about the bad weathers coming up, therefore providing relief to the related negative sentiment shocks. There can be various reasonings why it only happens for the cold end. One important guess is that the potential costs on avoiding cold weathers are much lower than that of avoiding hot weathers. For example, to comfort the negative sentiment brought by cold weather, one may just need to bring more clothes, in which case accurate forecasts are pretty useful. But to avoid heat, people may need to bear the cost of getting a new air condition in the first place, by which accurate instantaneous forecast is not very helpful. This hypothesis is in line with the global regression results that sentiment drop during heat is greater than during cold, and will be tested in the next section. Another explanation is that discomfort under misforecast heat may be taking longer than a day to reflect on people's sentiment. This would be partially tested in next subsection.

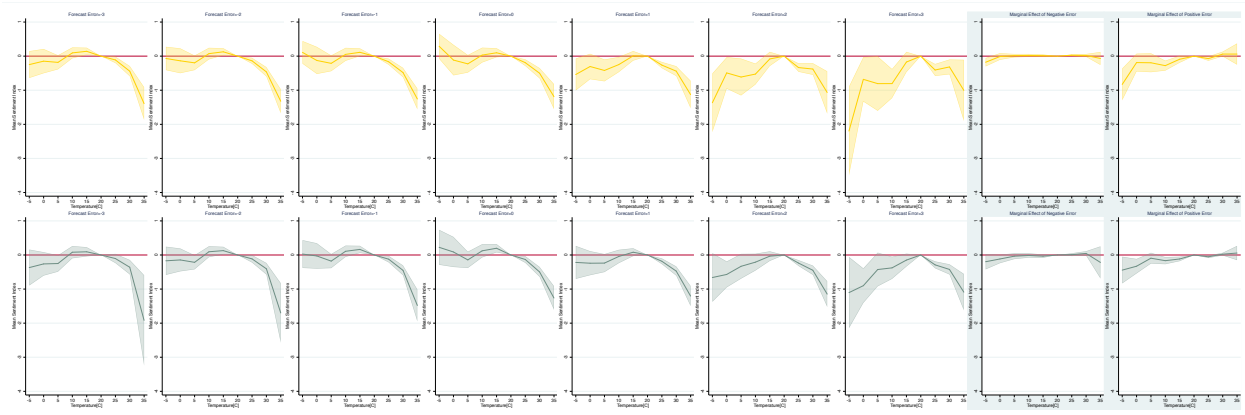
Another important observation comes comparing the size of the ME for positive and

negative forecast errors at the cold end. As shown in the plots, the negative effect of forecast errors is greater in magnitude for the positive rather than the negative forecast errors. Running F-test with null that the interactive estimates on the coldest bin $Tavg < -5C$ between negative and positive errors have been equal, the non-rationalized have F-statistics of 6.94 (p-value=0.0093), and the rationalized version has F-statistics of 1.41 (p-value=0.2377). Though the null is rejected at 1% for non-rationalized forecast errors, the rationalized version leads to no rejection of the null up to 10%. Similarly going through the first seven columns, positive forecast errors with the same magnitudes is seen related to greater size negative sentiment shocks of cold temperatures, comparing with negative forecast errors of the same size. In other words, people feel more unhappy about misforecast cold weathers whether the forecasts of the coldness are exaggerating rather than under-predicting. This is a bit surprising, as one may expect in real life the unhappiness would be greater when extreme cold is worse than expected. To explain, it may be related to the additional negative sentiments caused by over-alarming and over-preparation being set according to the forecasts, as they may generate greater cost than actually needed. Again, this hypothesis would be tested in the next subsection.

2.3.4 Lead and Lag Sentiment Responses

One of the additional question to ask is whether instantaneous forecast errors may relay to future sentiment responses, to real weathers or to future forecasts. As the trust towards a good forecasting system may be shaped and altered by these instantaneous forecast errors, the longer term impacts of these forecast errors are probable. To test that, I modify the

Figure 2.7: Baseline Interactive Regression Results, Original and Rationalized Forecasts



Note: Column (1)-(7): Left to right, $T^{real} - T^{forecast}$ increases from $-3C$ to $3C$; Column (8): Marginal effect of negative forecast error; Column (9): Marginal effect of positive forecast error; Top to bottom: Covariate $|T^{real} - T^{forecast}|$ non-rationalized (top) and rationalized (bottom); $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

main analysis Equation 2.1 under two conditions which I call the lead and the lag responses:

1. Lead: Suppose today people receive forecasts and realize the real weather, therefore realizing the forecast errors at time t as $T^{real} - T^{forecast}$, this error may affect people's sentiment response to the temperature forecasts for a future day to come $t + l$. This requires to run the regression of sentiment at time t to non-linear response of the forecast at $t + l$, with interaction of the instantaneous error at t:

$$
\begin{aligned}
Sentiment_{it} = \alpha &+ f_0(T^{forecast}_{it+l}) \\
&+ f_-(T^{forecast}_{it+l}) \times \mathbf{1}_{T^{real}_{it} - T^{forecast}_{it} < 0} \times |T^{real}_{it} - T^{forecast}_{it}| \\
&+ f_+(T^{forecast}_{it+l}) \times \mathbf{1}_{T^{real}_{it} - T^{forecast}_{it} \geq 0} \times |T^{real}_{it} - T^{forecast}_{it}| \\
&+ \gamma' \mathbf{X}_{it} + \epsilon_{it}
\end{aligned}
$$

Since reliable daily forecast is normally taken to be up to 3 days, maximum lead is

taken as $l = 3$.

2. Lag: Suppose some days ago people receive forecasts and realize the real weather, therefore realizing the forecast errors at time $t - l$ as $T^{real} - T^{forecast}$, this error may affect people's sentiment response to the real temperature days later at t. This require to run the regression of sentiment at time t to non-linear response of the forecast at t, with interaction of the previous forecast error at $t - l$:

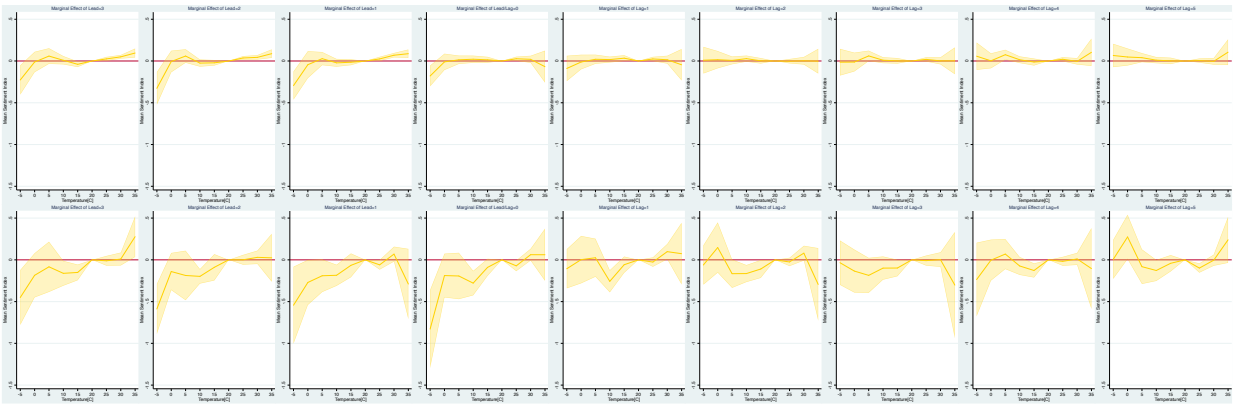
$$\begin{aligned} Sentiment_{it} = \alpha + f_0(T_{it}^{real}) \\ + f_-(T_{it}^{real}) \times \mathbf{1}_{T_{it-l}^{real} - T_{it-l}^{forecast} < 0} \times |T_{it-l}^{real} - T_{it-l}^{forecast}| \\ + f_+(T_{it}^{real}) \times \mathbf{1}_{T_{it-l}^{real} - T_{it-l}^{forecast} \geq 0} \times |T_{it-l}^{real} - T_{it-l}^{forecast}| \\ + \gamma' \mathbf{X}_{it} + \epsilon_{it} \end{aligned}$$

In this case I make the maximum lag response to previous historical forecast error at $l = 5$.

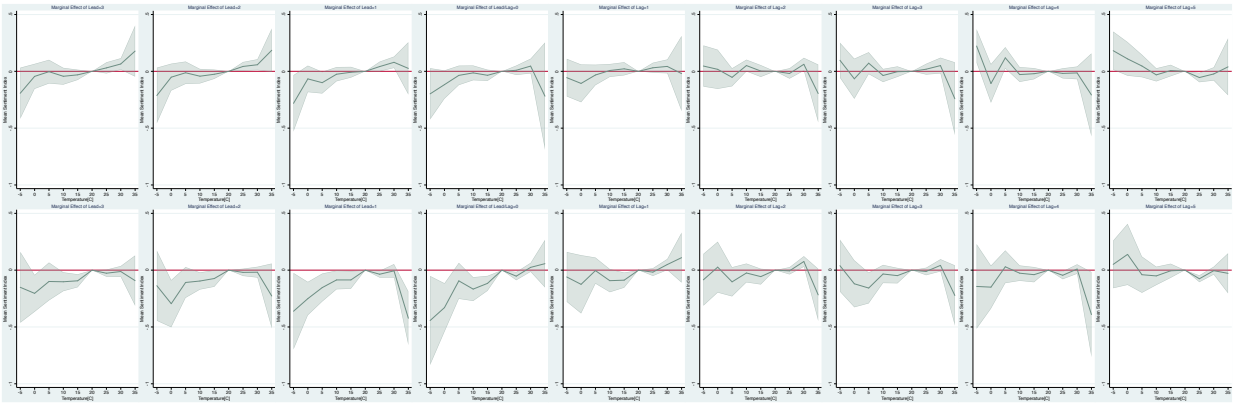
The results of these lead and lag responses are illustrated by the marginal effect plots for negative and positive errors separately, for both non-rationalized and rationalized forecasts, in Figure 2.8. Firstly looking at the cold end, negative ME persists for all leads with slightly smaller magnitude negative ME, for both negative and positive, non-rationalized and rationalized forecast errors. All the leads marginal effect curves are very similar to Column (4), which is the main regression. That is to say, people's sentiment response to future forecasts is similarly affected by instantaneous forecast errors realized today. However, as evidence by the similarity to the main results, this is likely to be the result of forecasts in the next 3 days are highly correlated with the real temperatures of today (correlation coefficient for $Tavg$ is

above 0.94 for all leads). On the contrary, the cold end marginal effect for the lag responses are almost all flat and statistically insignificant (occasionally small and positive), meaning that future cold temperature responses of sentiment is not quite related to the previous day forecast errors. There is not a delayed or enduring effect of the cold forecast errors on sentiment.

Figure 2.8: Interactive Regression with Leads and Lags of Forecast Error



(a) Non-Rationalized Forecast Errors



(b) Rationalized Forecast Errors

Note: Left to right: Marginal effect of absolute forecast error for leads $3, 2, 1$ days, no lead/lags, to lags $1, 2, 3, 4, 5$ days; Top to bottom: Negative and positive forecast error ME; $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

On the other hand unlike the main results, there exists instances in both leads and lags where hot end negative marginal effects exist with comparable magnitudes to the cold end (though statistical significance is mostly lacking except for rationalized forecasts with positive errors), especially for positive forecast errors. In summary, there may be impactful negative effects of current forecast errors on people's sentiment response to future heat forecasts up to 3 days, or previous forecast errors up to 4 days on current sentiment response to hot temperatures, especially when the forecasts are underestimates of likely hot days. On one hand, people may feel more unhappy about future's hot temperatures when previous forecast underestimates, because they may expect to see "hotter" than forecast temperatures. On the other hand, the extra negative sentiment shocks brought by a likely hot forecasts may add to unhappiness under heat in future days. This observation is more spurious however, because such negative sentiment seems only to be existing after 2 days of the errors realized.

Overall, this exercise implies that the impact of forecast errors can be delayed to future sentiment responses to forecasts or to real temperatures, but only likely for under hot temperatures when previous forecasts underestimate the heat. Lack of such evidence at the cold end suggests that people's sensitivity to inaccurate hot forecasts can be less instantaneous but more memorable, as larger underestimates today may spur people to expect even hotter and more unhappy temperatures tomorrow.

2.3.5 Response to Temperature Forecast Warnings

One of the argument for why sentiment responses at cold temperatures is more sensitive to the accuracy of forecasts when it is an overestimates of the coldness relies on the fact that over-preparation may cause additional unhappiness among the public. To verify this argument, I look at differentiated sentiment responses under cases when official heat or cold warnings has been issued by the CMA, guided by the 24-hour weather forecasts.

According to official definitions, I would define the "Orange" heat warning issued when 24-hour forecast $Tmax > 37C$, and "Red" heat warning when $Tmax > 40C$. The warnings issued for the cold side is more complicated, all must satisfy events of "considerable drop of temperatures in a short time frame within 24 hours" besides a requirement on the $Tmin$ forecasts. Since the "large temperature drop" event is hard to observe from my dataset (there are very few such instances comparing last day real temperatures and forecast temperatures in my dataset), I instead use a rough approximate for the "Orange" cold warning as $Tmin < -5C$ and "Red" cold warning as $Tmin < -10C$. Number of such instances is about the same as the heat warnings.

With those warning definitions, I run the following interactive regression:

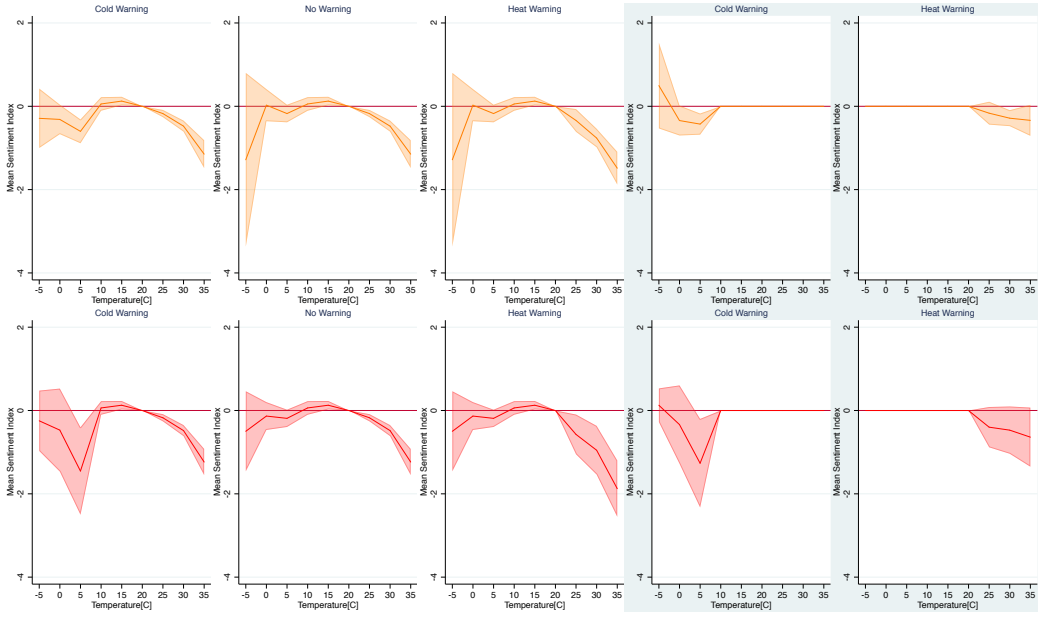
$$
\begin{aligned}
Sentiment_{it} = \alpha + f_0(T_{it}^{real}) \\
&+ f_{-}(T_{it}^{real}) \times \mathbf{1}_{Cold} \times \mathbf{1}_{ColdWarning} \\
&+ f_{+}(T_{it}^{real}) \times \mathbf{1}_{Hot} \times \mathbf{1}_{HeatWarning} \\
&+ \gamma' \mathbf{X}_{it} + \epsilon_{it}
\end{aligned}
$$

Here in this version of the regression, I allow asymmetrical interaction of the warning dummies as defined above, with heat warnings only interact on hot days and cold warnings only on cold days. With the non-linear functional form specified to be $5C$ bins as in the main, I let "Hot Days" to be of the hottest 3 bins and "Cold Days" the coldest 3 bins. The middle bins, including the reference bin, has no interactions with both warning covariates.

Results for sentiment responses to $Tavg$ has been shown in Figure 2.9. In summary, we see negative marginal effects of the cold warnings especially when actual temperature is not that cold (because when $Tavg < -5C$ the ME becomes positive and statistically insignificant). This can be interpreted as that when a cold warning is issued but coldness is not as severe, there is a greater sentiment drop than without the warning issued. This is consistent with the theory that over-reactions may affect people's negative sentiment more than unexpected extreme temperatures, as governmental temperature warnings are likely to accompany local policies to shut down public places and decrease transportations. Meanwhile, there is also a hot side negative marginal effect for the heat warnings. In fact, it shows a similarly symmetrical effect as the cold side, that when a heat warning is issued, people's

sentiment response to the heat is even greater when without the warning. This certifies somehow that extra costs exist with forecasts overstating the temperature extremes because forecasts based warnings can bring more inconvenience and unhappiness to people.

Figure 2.9: Interactive Regression with Cold and Heat Warnings



Note: Column (1)-(3): Left to right, Cold Warning, No Warning, Heat Warning; Column (4): Marginal effect of Cold Warning; Column (5): Marginal effect of Heat Warning; Top to bottom: Orange Warning and Red Warning; $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

2.4 Extensions and Robustness Checks

2.4.1 North vs South

In this part, I rerun the main analysis Equation 2.1 on separated subsample, 77 northern and 67 southern cities. The north-south region separation in China follows the historical convention of Qinling Mountain-Huai River, with which many policies differ across the bor-

der. For one example, northern cities are provided with central heating organized by the governments during cold months extending from October to April, but southern cities have to use their own installation of house heating system or through AC.

Figure 2.A.1 shows the result of this subsample analysis. First for the cold side, both north and south regions have negative contribution of forecast errors to the sentiment shock under extreme cold, though for the southern cities such effects only exist for positive rather than negative forecast errors. That is to say, southern sentiments are more relevant to forecasts when cold days are exaggerated rather than underestimated. Also noted from the first seven columns of the plots, the negative sentiment shocks for southern cities are in general greater than that for the northern cities by magnitudes. That is likely due to better adaptation to cold weathers for the colder cities in the north. For the hot end, there is some negative ME with considerable magnitudes for positive non-rationalized errors, and negative rationalized errors. These evidence, though inconclusive because of inconsistency between rationalization and less statistical significance, still may show the sensitivity of sentiments under heat to forecast errors only persists among people living under warmer climates. Likewise, the negative sentiment shocks under heat for northern and colder cities are mostly in magnitude greater than those in the southern part of the country.

These results, comparing with the main, may imply that the role of forecast errors depend on whether long run climate is warm or cold. However, the conclusion is rather inconclusive and it is unclear whether this is more related to the climates or policies such as the accessibility of central heating. Instead, I would summarize the results of this exercise as the

relative consistent cold end marginal effects of forecast error regardless of south and north.

2.4.2 Low vs High Income

In Figure 2.A.2 it shows the interactive regression result for cities separated by GDP per capita into two groups, the low versus the high, with GDP per capita cut-off defined by 2014 median among the 144 sample cities. From that plot, the cold side responses again looks quite similar across different income group, with negative marginal effects of forecast errors especially for positive errors. The hot side shows some evidence for both non-rationalized and rationalized (but statistical significance only to rationalized) that negative marginal effect for sentiment exists, either the forecast error is positive or negative, only for the high income subsample. These observations then provide a partial proof for the argument that the avoidance cost of cold is smaller versus hot. Because of the relative higher cost of heat avoidance when facing a hot weather forecast (e.g., obtaining an AC), smaller negative sentiment shocks come after an accurate hot weather forecasts to those being able to pay the higher cost in the first place, i.e., those residing in higher income cities. For those without the resource to pay for heat avoidance in the first place, their negative sentiment shocks would be less relevant to forecast accuracy. On the other hand, because cold avoidance is more affordable for all income levels, the marginal effect of forecast error is similarly negative for both income groups.

2.4.3 Small vs Large Long Run Forecast Errors

In real life, people likely have different trusts for daily weather forecasts based on their longer run reliability. To access that, I estimate the city-by-year root-mean-squared-error (RMSE) of daily temperature forecasts in 2014, and divide the 144 cities in the sample into low and high RMSE cities with median as the cut-off. The subsample analysis is run In Figure 2.A.3. Again, it shows that cold side marginal effects of the instantaneous forecast errors are similarly negative across different cities with different long run forecast accuracies. Meanwhile, the hot end negative marginal effects only exist in the group where RMSE is low, or say forecasts are generally more accurate. In other words, people living in cities likely trusting more about daily temperature forecasts because of better long run performance tends to experience less unhappiness during hot days because of smaller instantaneous forecast errors (especially when forecasts are overestimates), because for them those would likely be heat shocks that are more unexpected. These again verifies the theory that sentiment responses related to heat is influenced by the fact that heat avoidance may occur with higher cost. For example, when hot days are forecast and preparation made accordingly only when forecasts are trusted, a coming forecast error would contribute to greater sentiment drops. Meanwhile on the cold side, the cost of avoidance is low enough that people may take precautions whether or not they trust the forecasts, and their sentiment improves if the instantaneous forecast errors are realized to be smaller.

2.4.4 Holidays vs Working Days

Figure 2.A.4 runs the main interactive regression for subsamples separated by time instead of by space, into holidays (including weekends) or working days. Likewise, the cold side looks quite similar across the two subsamples, with negative marginal effects of forecast errors especially for positive ones. However, the hot side only has negative marginal effects with considerable magnitude and statistical significance for holidays rather than working days. It means that during holidays people's unhappiness during heat is more significant when forecast fails to predict it. I think this again can be partially explained by the possible higher cost of heat avoidance. For instance, during holidays people are more likely to go outdoors and have more flexibility to adjust their activities according to forecasts, thus a hot day forecast can change people's plan during holidays. When the real temperature is realized, it would be too late and too costly to rearrange if forecast errors are high. As a result, sentiment drop is larger. Meanwhile during working days when such change of schedule is not that flexible in the first place, the accuracy of hot day forecasts will be less relevant to the negative sentiment shocks displayed during heats.

2.4.5 Responses to Tmin and Tmax

One argument regarding beliefs formed on weather forecasts is that people may be taken the direct information in weather broadcasting, rather than the inferred information. In the setting of this paper, the weather forecast program in China broadcasts $Tmin$ and $Tmax$, while $Tavg$ has to be inferred as their mean. As a result, people's response may not be very

relevant to the forecast of $Tavg$, but instead to $Tmax$ and $Tmin$. The response to $Tmin$ has the additional problem that it is usually observed during midnight, when social media activities are the lowest. On the other hand, response to $Tmax$ may be less representative during the cold days as the sentiment index is likely reflecting on worse weather scenario of the day.

To verify, I rerun the main regression with forecast error defined for $Tmin$ and $Tmax$ respectively as the covariates (and the response of sentiment too), results shown in Figure 2.B.1. In general, the hot end result persists for both, that the marginal effect on forecast errors for non-rationalized and rationalized, negative and positive errors are mostly close to zero and statistically insignificant. The hot end marginal effect for $Tmax$ even flips to positive significant at $Tmax \geq 35C$, meaning that greater the forecast error for daily maximum temperature the better the sentiments. However, since the negative sentiment response to hot days are always sharp and persistent, the relative small size of those positive ME has been quite insignificant in reducing the negative heat-related sentiment drop towards zero.

On the other hand, the cold end results for $Tmin$ are only negative and significant when forecasts has positive error (i.e., daily minimum temperature forecast is exaggerated). Meanwhile, the negative ME for $Tmax$ has been similar (and F-test not rejected for both non-rationalized and rationalized) for both positive and negative errors. Comparing the MEs of positive errors for $Tmin$ and $Tmax$, their magnitudes have been quite similar at the corresponding coldest bins. Those combines to bring the negative ME for $Tavg$ of greater magnitude for positive forecast errors. In implications, people's greater unhappiness during

cold weathers are driven by both exaggerated $Tmin$ and $Tmax$ forecasts, while undermined $Tmin$ forecasts seems not to be playing an important part. Overall, it verifies the guess that $Tmin$ response is less relevant to the sentiment index because most posts on social media are due in daytime. However, positive forecast errors with $Tmin$ still matters by causing over-reactions to the extreme cold, possibly through cold warning issued by the government.

2.4.6 Different Functional Forms

Besides the non-parametric bin regression used as f in Equation 2.1, I also run the analysis with two of the more parametric options, polynomial order 3 (cubic polynomial approximation) and restricted cubic splines with knots $(0, 5, 10, 15, 20, 25)$ for $T = Tavg$, for the purpose of robustness checks. The results are shown in Figure 2.B.2. Comparing with the middle row bin regression, both parametric functional forms display smoothing in the extreme temperature ranges such that the negative marginal effect of forecast errors at the cold end has been significantly reduced both in magnitude and in statistical significance, for both non-rationalized and rationalized forecasts. Actually, the significant negative sentiment shock on the cold end has been almost disappearing for both parametric approximations. In fact, the results at cold end with the non-parametric bin regression likely have been driven by only 0.5% of the observation that $Tavg < -5C$ (about the same percentage is at the hot end, $Tavg \geq 30C$). To robustness check whether the small number of outliers may be the only driver to the main result, I will run the next subsection with trimming.

2.4.7 Trimming Extremal Temperatures or Forecast Errors

In Figure 2.B.3, I run the main interactive regression with trimming of sample. The top panel shows trimming of extreme 1% cold and 1% hot temperature days. As a result, though the extreme temperature bins are deleted from the main results, the analysis still shows the same results of negative marginal effects of instantaneous forecast error $|T^{real} - T^{forecast}|$ on sentiment under cold temperatures $Tavg < 0C$, for both positive and negative errors. It also pertains the same results that positive forecast errors have greater marginal effect on sentiment than negative ones after the trimming.

In the bottom panel, I run the analysis with trimming of top and bottom 1% forecast error, i.e., trimming the very negative and very positive instantaneous forecast error $T^{real} - T^{forecast}$. These will mainly trim days when the forecasting system may occasionally do a bad job but the error would not necessarily persist. Again, the plots are very much similar to the whole sample in main analysis, affirming that the main results are not driven by extremal forecast errors.

2.4.8 Interactive Regression with Long Run Forecast Errors

People's sentiment response may not only relying on instantaneous forecast errors, but also on long run forecast accuracy, as previously explored. In this part, I rerun the interactive regression in the main design but with a single-sided covariate for linear interaction, the long run temperature forecast root-mean-squared-error (RMSE) by city-year. Correspond-

ing results are shown in Figure 2.C.1. From those, we can reach a similar conclusion as the instantaneous forecast error, that greater long-run average forecast errors enhance people's negative sentiment shocks during cold, but not during hot. In explanations, more accurate forecasts (with lower long run RMSE) would allow people to prepare more promptly during coming cold weathers, thus relieving the negative sentiment shocks after real temperature is realized. However, similar effect could not happen for the hot days since there is greater cost to avoid heat, thus even with accurate forecasts effective actions cannot be made in time enough to offset the coming negative sentiment shocks. The magnitudes of these marginal effects at cold end has been of same order of magnitudes with the instantaneous forecast errors, stating that people's sensitivities to forecast errors at long or short runs have been quite similar in signs and sizes.

2.4.9 Naive Forecasts

In this paper, the forecasts people care about are assumed to be the "professional" forecasts provided by CMA. However, people may also be taking their own expectations of tomorrow's weathers based on historical climates. One question would be whether the weather forecasting system, invested and managed by the government and aiming to provide higher quality forecasts to the general public than people's "naive" expectations, would indeed impact their sentiment responses rather than the naive forecasts.

To disentangle between the effects of professional and naive forecasts, I consider a self-prediction model for 24-hour temperature forecasts based on historical real temperatures

using a autoregressive model AR(7):

$$T_{is} = \alpha_i^t + \sum_{k=1}^{7} \beta_{ik}^t T_{is-k} + \phi_{iw}^t + \delta_{im}^t + \gamma_{iy}^t + \epsilon_{it}^t$$

This is run for every city i and every day t with rolling windows $s = s_0, s_0 + 1, ..., t - 1$, where s_0 is January 1st, 2013. Fixed effects are for weekday (ϕ_{iw}), month (δ_{im}) and year (γ_{iy}).

After running these AR(7) regressions, I compile the "naive forecasts" indexed by city i date t based on the series of estimated coefficients, $\hat{T}_{it}$. These naive forecasts are 97% correlated with the real temperatures. Compare the average naive forecast absolute errors within the sample with the state professional forecasts, the naive forecasts features $0.7C$ lower prediction errors than the non-rationalized state forecasts, but up to $0.3C$ higher errors than the rationalized state forecasts. In summary, the naive forecast correct some of the negative bias persists over the non-rationalized forecasts, but would still be less accurate than if individual rationalized that raw forecast information.

With naive forecast $\hat{T}_{it}$, interactive regressions are repeated in Figure 2.C.2. Overall, the cold side still have some statistically significant negative marginal effects of both positive and negative forecast errors at the cold end, but it is rather unstable (for negative errors, ME only persists for the coldest bin $Tavg < -5C$, for positive error, it goes negative before the coldest bin but switch to positive when $Tavg < -5C$) and is with a smaller magnitude even comparing with the less accurate non-rationalized state forecasts. What is interesting is on the hot side, which shows a significantly negative marginal effect under heat only for

positive naive forecast errors. This has not been seen from the main results. In interpretation, if the naive forecast based on historical expectation tells an underestimate of the heat, unhappiness generated during heat is greater. Though the size of this marginal effect can be seen quite small comparing with the large sentiment drop on the hot end in the first seven columns, its estimate has been statistically significant and comparable to the ME at the cold end.

This exercise has suggested two points. Firstly, the ease of naive temperature estimations cannot replace the professional forecasts provided by state agency. The non-rationalized professional forecasts are less accurate, but the general public still seems to take them seriously as evidence by the stronger marginal effects of cold temperature forecast errors. However, it also implies that naive expectations can matter to people's sentiment response to hot weathers. It seems that such expectation will have its effect on reducing the negative sentiment shock encountering heat if the expectation is realized to be accurate enough. Therefore, the national forecasting system is important in shaping people's sentiment responses to cold temperatures, so is likely contributing to the social welfare in China. But meanwhile, people's sentiment is also linked to more naive expectations based on historical weathers, especially during hot days. As the state forecast models also take historical weathers as inputs and how people exactly form their naive expectations are unknown (and unlikely to be known except running field experiments with surveys), this conclusion for the reliance on this naive forecast is only suggestive and can also be just part of the results with state forecasts.

2.4.10 Sentiment Response to Precipitations

Besides temperatures, sentiments are also likely affected by precipitation events, such as rain, snow, thunderstorms. As a result, the accuracy of precipitation forecasts can be of importance to people's sentiment expressed on social media. For example, people would complain if it is sunny but forecast to be rainy, even more than to note that today's temperature is $30C$ rather than $29C$ as forecast.

The major barrier to study is the different formats of precipitation records in forecast and real weathers. For forecasts, categorized precipitation events are reported, while for real weather records it is the actual precipitation in mm. Both methods of recording have their limitations, with the categorical forecasts likely not representative for all area covered by a city, and the numerical records notably biased on due to rainfall collection errors per weather station as well as through approximation errors in the ERA-Interim modeling process. Nevertheless, all these has made it hard to match between forecast and real precipitations, and to compile a reliable metric for the precipitation forecast accuracy. In this exercise, I would take a simplified approach by narrowing down precipitations into two broad categories, the "Good Weather" (forecasts Sunny/Cloudy/Shady or non-positive real precipitations), and the "Bad Weather" (forecasts otherwise with Rain/Snow/Fog or positive real precipitations). Then I take the accuracy metric of the precipitation forecasts as the average probability of "Good/Bad Weather" being forecast incorrectly.

I run the interactive regression with an empirical design similar to Shrader (2020):

$$Sentiment_{it} = \alpha + \beta BadWeather_{it} + \gamma BadWeather_{it} \times Error_{it} + \delta' \mathbf{X}_{it} + \epsilon_{it} \tag{2.3}$$

Where $BadWeather$ represents whether city i in day t has a real positive precipitation being notated as "Bad Weather". $Error_{it}$ is the precipitation forecast error as defined above, I will take multiple frequencies besides instantaneous daily error of the dummy 1 or 0, at city-year or city-month average level. Again, I control for city and day FE, then cluster standard errors by city.

Table 2.C.1 then presents the results. Overall, a day in "Bad Weather" with non-zero precipitations generate as large as 0.3 decrease in sentiment index, comparable to the average negative shock under cold but much smaller than the negative shock under hot (Figure **??**). Estimates across the columns have been statistically significant at 1%, showing the robustness of such effects as expected by intuition. However, the interactive coefficients of "Bad Weather" with precipitation forecast errors have been consistently positive with a similar magnitude as the direct estimates, especially statistically significant when taking the error to be instantaneous. In another word, the negative sentiment shock to "Bad Weather" would be reduced significantly when this precipitation event is not predicted and instead weather is forecast to be "Good". This is quite counterintuitive because one may expect people's mood to drop when precipitation hits unexpectedly, while it seems that erroneous forecasts actually help to improve the sentiment. Though I may blame the fact that forecasts for non-capital cities are approximated by their capitals, the same regression with capital cities

only (Table 2.C.2) shows the very similar results. I may still argue the reliability of these results because of the inaccuracy of precipitation records, or my simplification of "Good/Bad Weather", but by far the overall results for sentiment response to precipitation forecasts have shown the contrary of better forecasts helping with people's sentiment during bad weathers, as proved in the main results with sentiment-temperature responses. However, it might be another instance showing that bad weather warnings add to the negative sentiment of people on social media, even more than the negative sentiment shock caused by precipitation weathers themselves.

2.5 Conclusion

In this paper, I analyze the effect of instantaneous temperature forecast errors on the social media sentiment responses to temperatures in China. For both non-rationalized and rationalized forecast beliefs, there are negative significant marginal effects of daily temperature forecast errors during cold weathers, regardless of positive or negative errors. Meanwhile on the hot end, the negative sentiment shock in response to heat has always been sharp and significant regardless of the sizes or signs of these instantaneous forecast errors. In short, people are more unhappy about misforecast cold weathers but feel equivalently unhappy during misforecast or correctly forecast hot weathers. In general, these results match the intuition that better forecasts can better prepare people about the bad weathers coming up, hence providing relief to the coming negative sentiment shocks. In addition, the negative marginal effects of forecast errors are greater in magnitude for positive comparing with the

negative forecast errors, meaning that exaggerations of cold is worse to people's sentiment than underestimates.

Further analysis also reveal reasonings of these results. For greater negative sentiment responses when cold forecast is an exaggeration, it can be related to the additional negative shocks brought by cold warning messages issued according to the forecasts. For small sentiment sensitivity to forecast errors on the hot end, it may be that there takes more than one day to experience the negative effect of hot temperature forecast errors. Otherwise, it is most likely that the relatively low cost of cold avoidance comparing with heat avoidance limits the sensitivity to hot forecast errors only to subsamples of higher income, greater long run forecast accuracy, and during holidays.

Overall, these results provide evidence for the positive welfare impacts of improving weather forecasting technology. Maximum improvement in sentiment per $1C$ decrease of daily forecast error for a cold day is at maximum 0.8 when daily average temperature is below $-5C$, 1.4% of the average median sentiment index across days under this cold temperature. This value is also comparable to the negative impact on sentiment index per one unit increase of pollutant in Zheng etal. (2019). Comparing with the increase of average median day-of-week sentiment from Tuesday to Saturday, this marginal effect is as large as 42% of the weekday-weekend contrast. Coarsely related to GDP per capita of a city in 2014 by a simple OLS between their median sentiment and income level, such marginal effect accounts for the cost of 726 Yuan (currency in 2015) per $1C$ forecast error per day. Considering there are on average 17.2 days in 2014 when $Tavg < -5C$ across the 144 cities in sample, this

means up to 20% the average city GDP per capita lost if this marginal $1C$ forecast inaccuracy carries persistently for a whole year.

This project serves as an analysis of the welfare impacts of improving weather forecast accuracies in modern days. The results provide initial evidence for the value of accurate weather forecasts and the relatively high economic returns of investing in a modern weather forecasting system that create large social benefits for billions of people in China. For future explorations, more research need to be done in verifying the detailed mechanism through which forecast errors has been realized and impacting individual subjective well-beings. For example, forecast errors can directly impact people's sentiment or health, or it can affect how well people prepare for extreme weathers, hence affecting people's choice in labor and transportation. To start with, breakdown analysis by occupations, genders and microblog topics of discussion may reveal more. In real life, social media is also a tool for exchanging information about extremal weather events. Therefore, more analysis regarding the microblog texts may further reveal the process of shaping beliefs and forming sentiment responses around forecast information.

2.6 References

Allen, Roy, J. Graff Zivin, and Jeffrey Shrader. "Forecasting in the presence of expectations." *The European Physical Journal Special Topics* 225, no. 3 (2016): 539-550.

Bruni, Luigino. "The technology of happiness and the tradition of economic science." *Journal of the History of Economic Thought* 26, no. 1 (2004): 19-44.

Carleton, Tamma A., Amir Jina, Michael T. Delgado, Michael Greenstone, Trevor Houser, Solomon M. Hsiang, Andrew Hultgren et al. *Valuing the global mortality consequences of climate change accounting for adaptation costs and benefits*. No. w27599. National Bureau of Economic Research, 2020.

Coen, Deborah R. "A brief history of usable climate science." *Climatic Change* 167, no. 3 (2021): 1-17.

Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter." *PloS one* 6, no. 12 (2011): e26752.

Downey, Mitch, Nelson Lind, and Jeffrey G. Shrader. *Adjusting to Rain Before It Falls*. Working Paper, 2021.

Easterlin, Richard A. "The economics of happiness." *Daedalus* 133, no. 2 (2004): 26-33.

Edwards, Paul N. "History of climate modeling." *Wiley Interdisciplinary Reviews: Climate Change* 2, no. 1 (2011): 128-139.

Fox, Glenn, Jason Turner, and Terry Gillespie. "The value of precipitation forecast information in winter wheat production." *Agricultural and Forest Meteorology* 95, no. 2 (1999): 99-111.

Glatzer, Wolfgang, Laura Camfield, Valerie Moller, and Mariano Rojas. "Global handbook of quality of life." *Exploration of well-being of nations and continents* (2015).

Hsiang, Solomon. "Climate econometrics." *Annual Review of Resource Economics* 8 (2016): 43-75.

Kahneman, Daniel, and Alan B. Krueger. "Developments in the measurement of subjective well-being." *Journal of Economic perspectives* 20, no. 1 (2006): 3-24.

Kryvasheyeu, Yury, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. "Rapid assessment of disaster damage using social media activity." *Science advances* 2, no. 3 (2016): e1500779.

Lazo, Jeffrey K., Rebecca E. Morss, and Julie L. Demuth. "300 billion served: Sources, perceptions, uses, and values of weather forecasts." *Bulletin of the American Meteorological Society* 90, no. 6 (2009): 785-798.

Martinez, Andrew B. "Forecast Accuracy Matters for Hurricane Damage." *Econometrics* 8, no. 2 (2020): 18.

Mincer, Jacob A., and Victor Zarnowitz. "The evaluation of economic forecasts." In

Economic forecasts and expectations: Analysis of forecasting behavior and performance, pp. 3-46. NBER, 1969.

Nurmi, Vaino, Adriaan Perrels, Pertti Nurmi, Silas Michaelides, Spyros Athanasatos, and Matheos Papadakis. "Economic value of weather forecasts on transportation: Impacts of weather forecast quality developments to the economic effects of severe weather." *EWENT FP7 project* (2012).

Rosenzweig, Mark R., and Christopher R. Udry. *Assessing the Benefits of Long-Run Weather Forecasting for the Rural Poor: Farmer Investments and Worker Migration in a Dynamic Equilibrium Model.* No. w25894. National Bureau of Economic Research, 2019.

Shrader, Jeffrey. "Expectations and adaptation to environmental risks." *Available at SSRN 3212073* (2020).

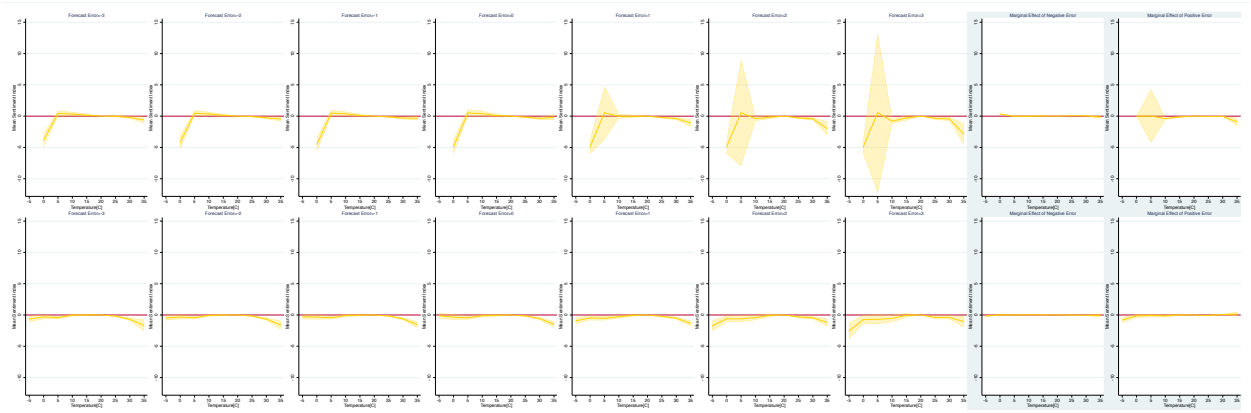
Wang, Jianghao, Nick Obradovich, and Siqi Zheng. "A 43-million-person investigation into weather and expressed sentiment in a changing climate." *One Earth* 2, no. 6 (2020): 568-577.

Yuan, Huiling, Min Sun, and Yuan Wang. "Assessment of the benefits of the Chinese Public Weather Service." *Meteorological Applications* 23, no. 1 (2016): 132-139.

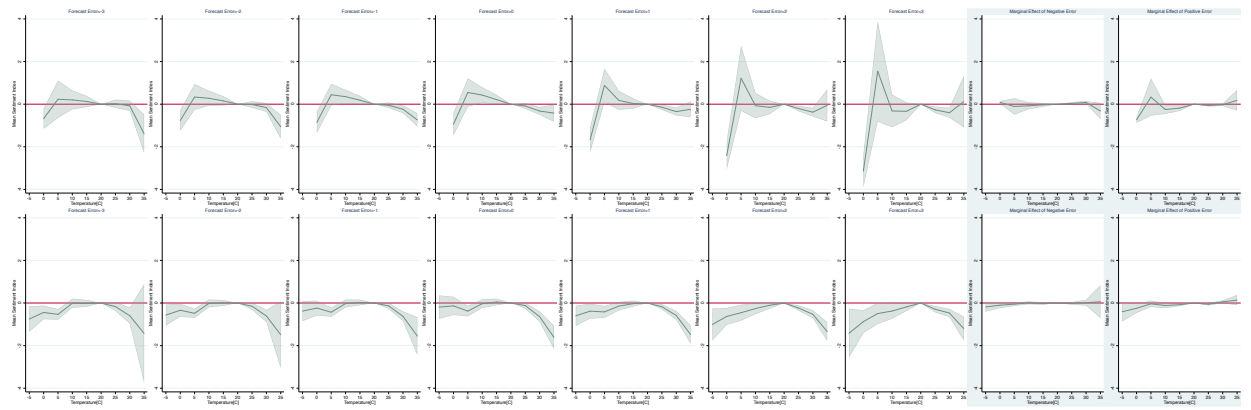
Zheng, Siqi, Jianghao Wang, Cong Sun, Xiaonan Zhang, and Matthew E. Kahn. "Air pollution lowers Chinese urbanites' expressed happiness on social media." *Nature human behaviour* 3, no. 3 (2019): 237-243.

2.A Separate Subsamples Analysis

Figure 2.A.1: Interactive Regression with North-South Separation



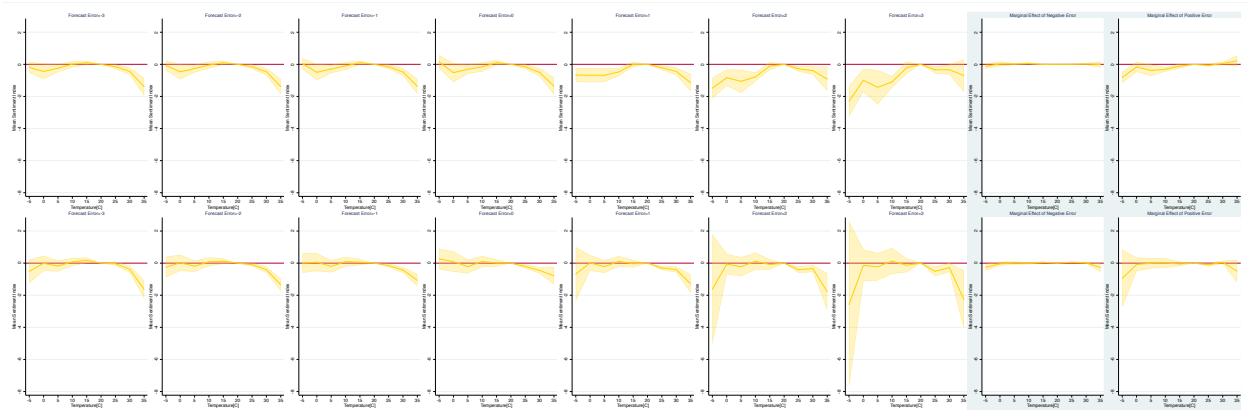
(a) Non-Rationalized Forecast Errors



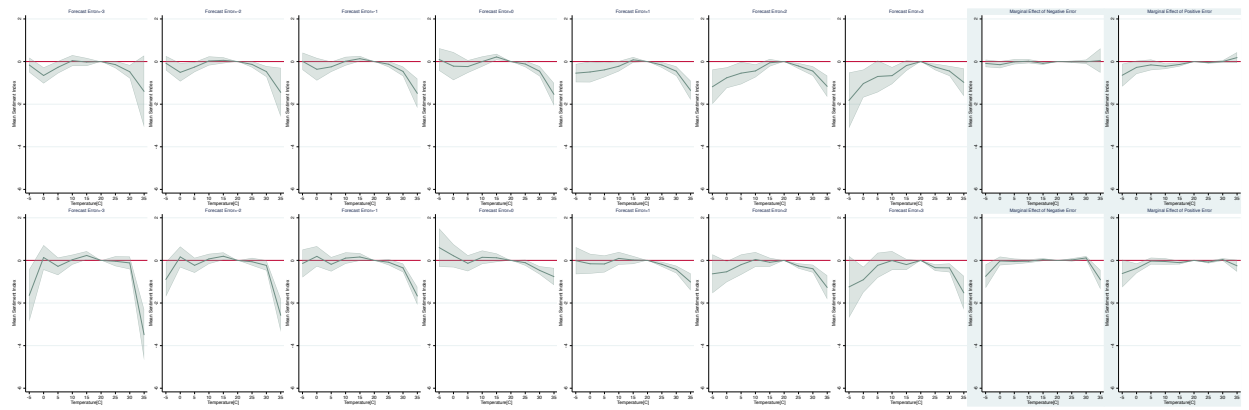
(b) Rationalized Forecast Errors

Note: Column (1)-(7): Left to right, instantaneous forecast error $T^{real} - T^{forecast}$ increases from $-3C$ to $3C$; Column (8): Marginal effect of negative forecast error; Column (9): Marginal effect of positive forecast error; Top to bottom: Southern cities (67) VS northern cities (77); $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

Figure 2.A.2: Interactive Regression with Low VS High Income



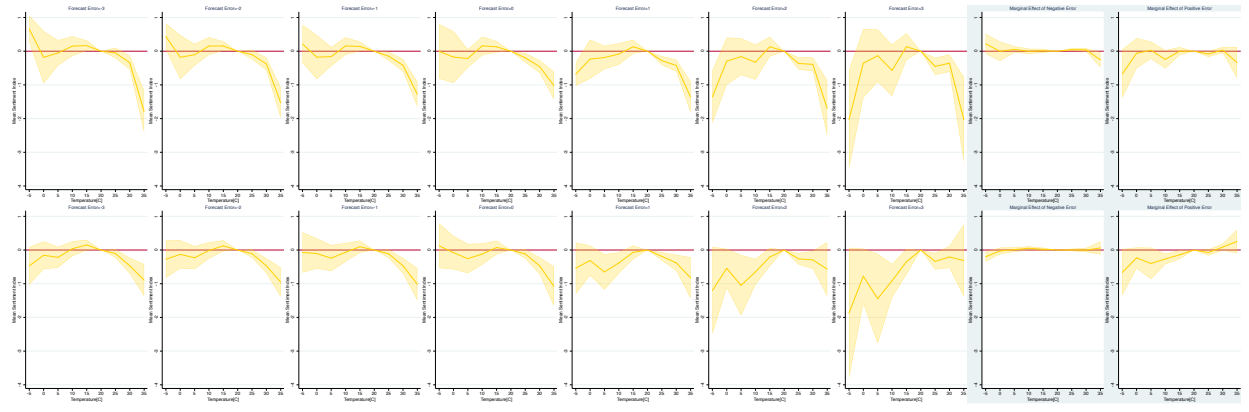
(a) Non-Rationalized Forecast Errors



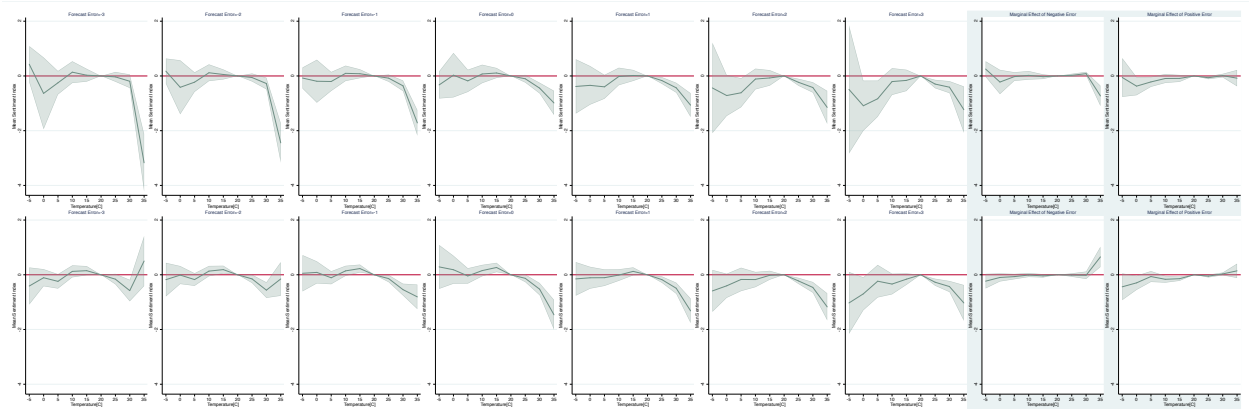
(b) Rationalized Forecast Errors

Note: Column (1)-(7): Left to right, instantaneous forecast error $T^{real} - T^{forecast}$ increases from $-3C$ to $3C$; Column (8): Marginal effect of negative forecast error; Column (9): Marginal effect of positive forecast error; Top to bottom: Low Income VS High Income, cut-off by GDP per capita in 2014 by median over the 144 cities; $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

Figure 2.A.3: Interactive Regression with Low VS High Long-Run Average Forecast Errors



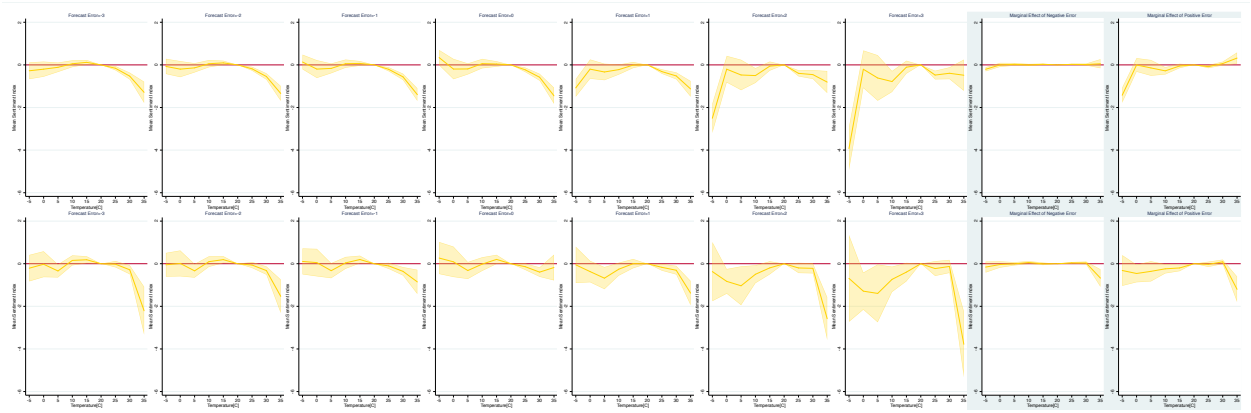
(a) Non-Rationalized Forecast Errors



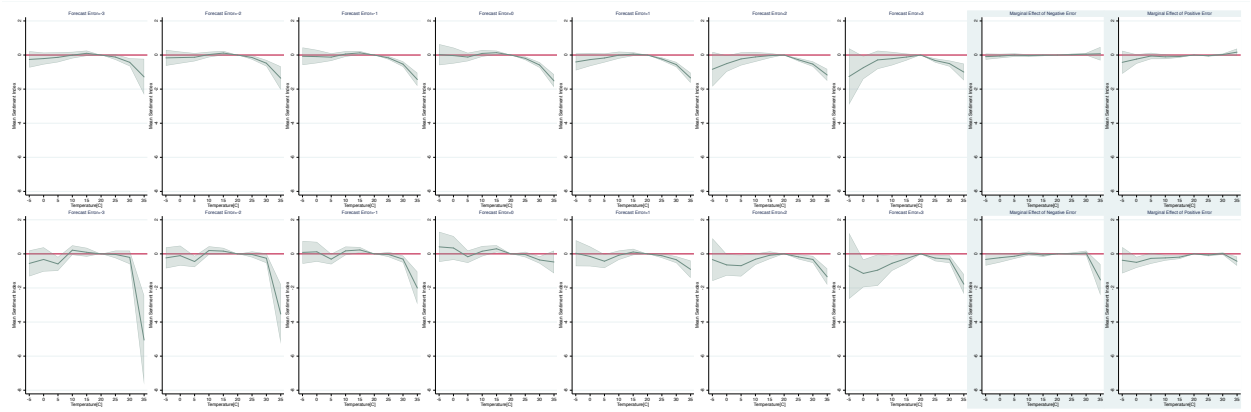
(b) Rationalized Forecast Errors

Note: Column (1)-(7): Left to right, instantaneous forecast error $T^{real} - T^{forecast}$ increases from $-3C$ to $3C$; Column (8): Marginal effect of negative forecast error; Column (9): Marginal effect of positive forecast error; Top to bottom: Low VS High Yearly Forecast $RMSE$, cut-off by median over the 144 cities; T = $Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

Figure 2.A.4: Interactive Regression with Working VS Non-Working Days



(a) Non-Rationalized Forecast Errors

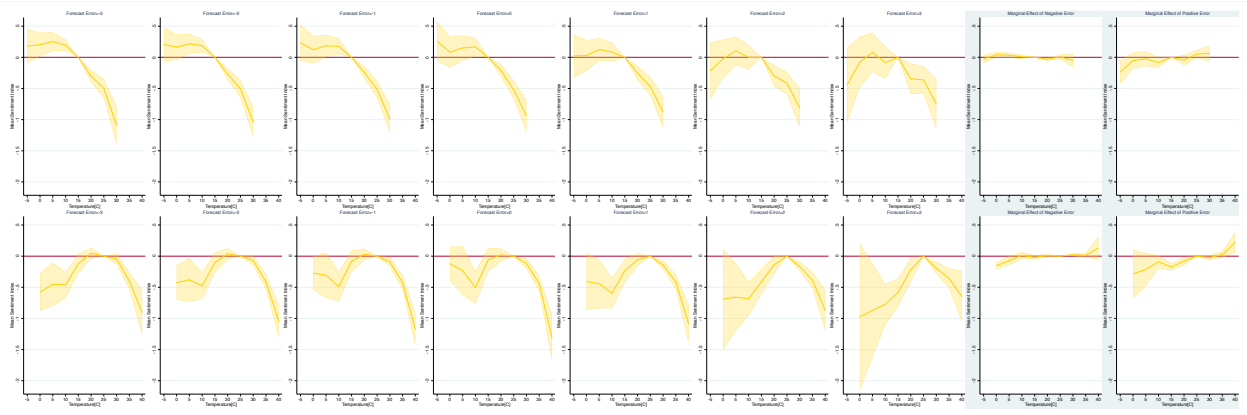


(b) Rationalized Forecast Errors

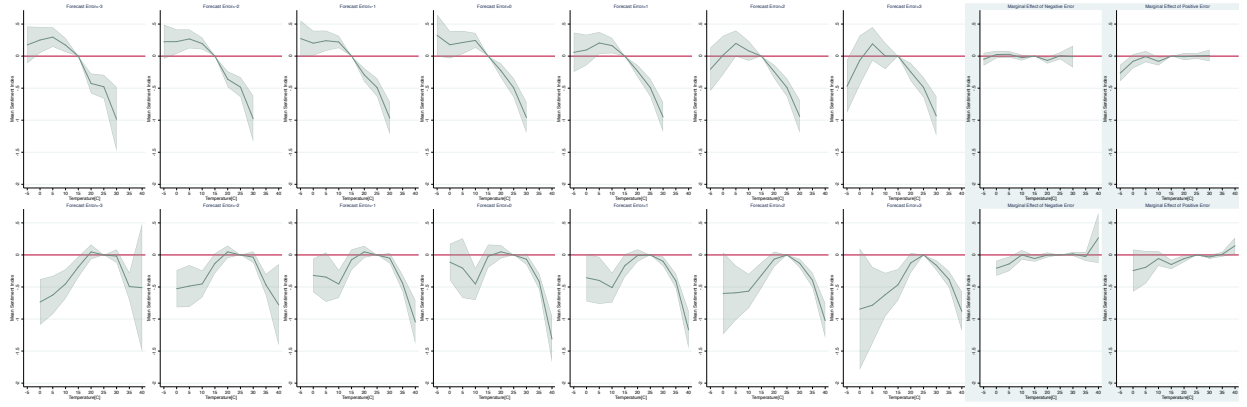
Note: Column (1)-(7): Left to right, instantaneous forecast error $T^{real} - T^{forecast}$ increases from $-3C$ to $3C$; Column (8): Marginal effect of negative forecast error; Column (9): Marginal effect of positive forecast error; Top to bottom: Working Days VS Holidays and Weekends; $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

2.B Robustness Checks

Figure 2.B.1: Interactive Regression with Tmin and Tmax



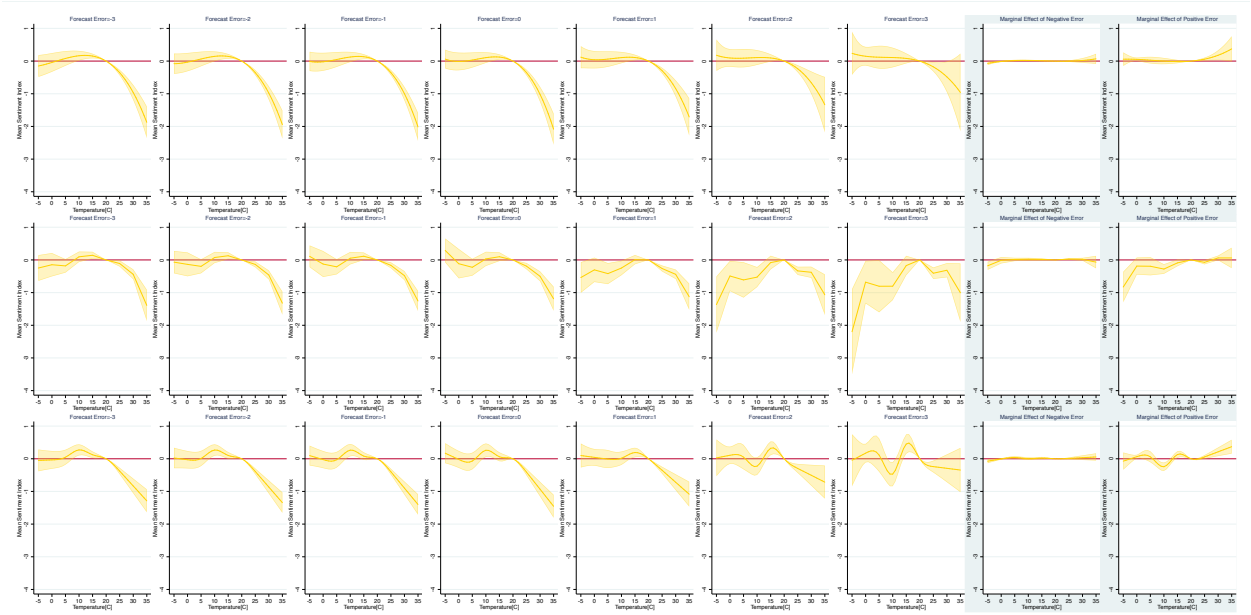
(a) Non-Rationalized Forecast Errors



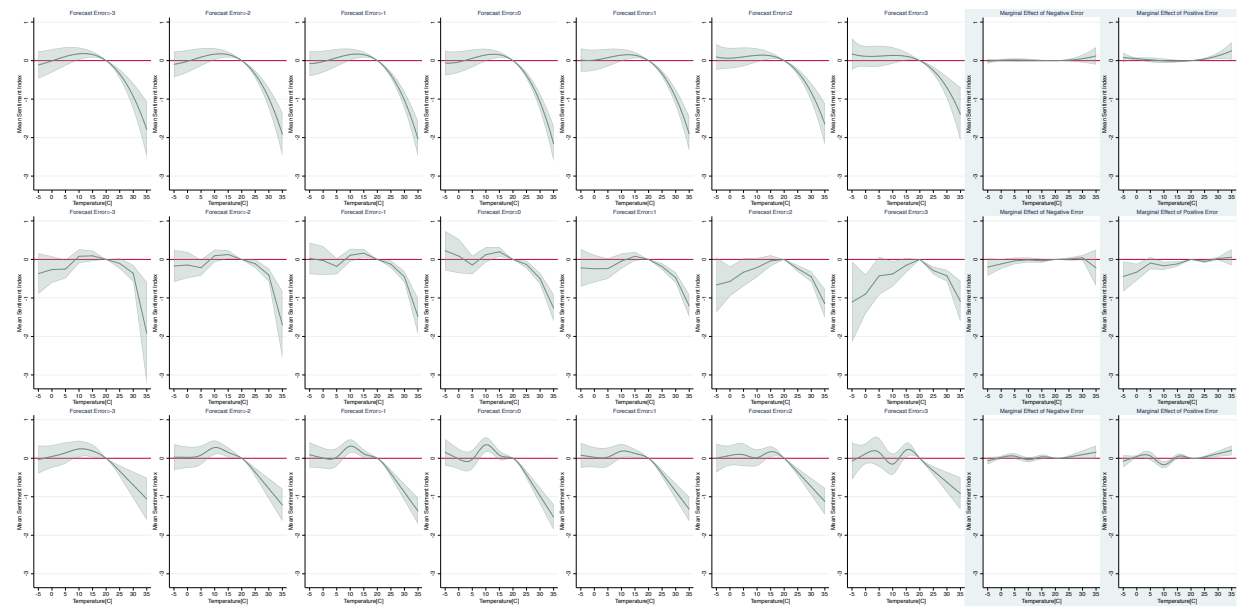
(b) Rationalized Forecast Errors

Note: Column (1)-(7): Left to right, instantaneous forecast error $T^{real} - T^{forecast}$ increases from $-3C$ to $3C$; Column (8): Marginal effect of negative forecast error; Column (9): Marginal effect of positive forecast error; Top to bottom: $T = Tmin, Tmax$, reference bin $[10C, 15C)$ and $[20C, 25C)$ respectively; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

Figure 2.B.2: Interactive Regression with Different Functional Forms



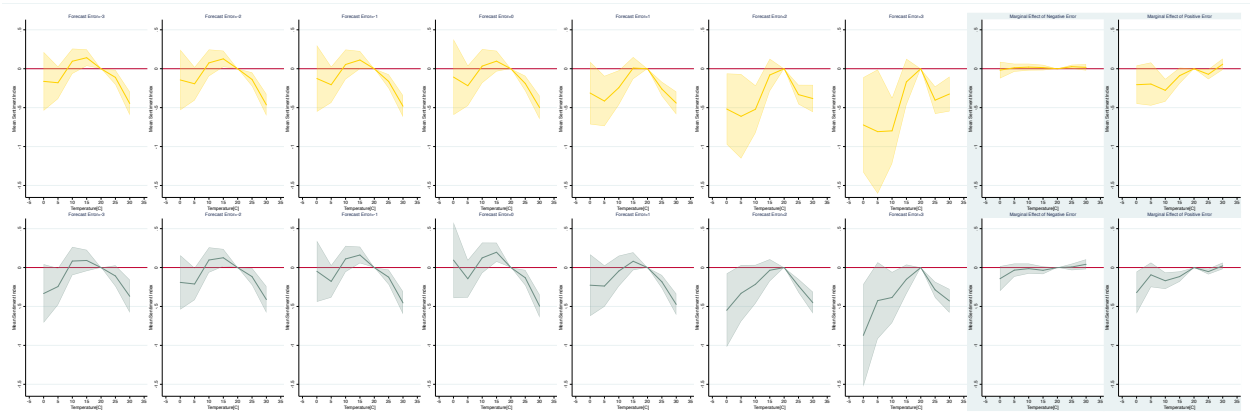
(a) Non-Rationalized Forecast Errors



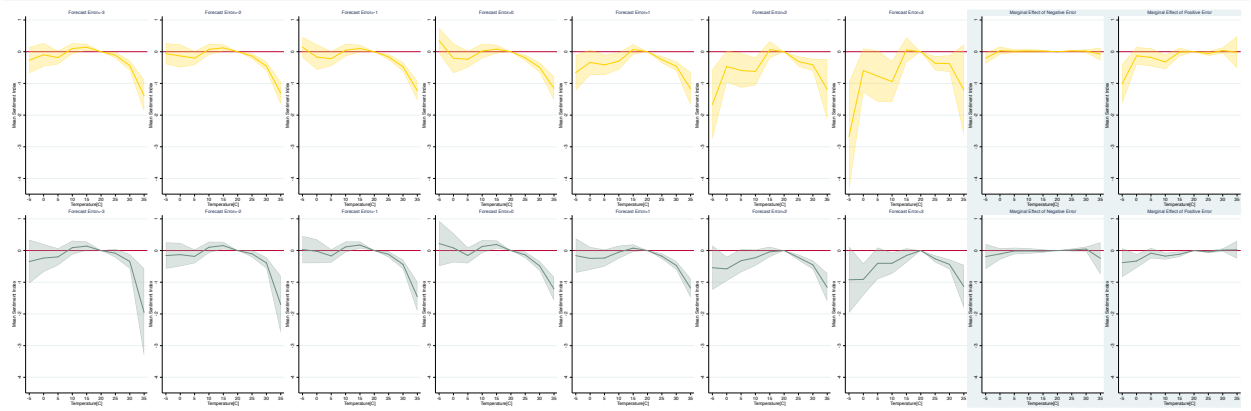
(b) Rationalized Forecast Errors

Note: Column (1)-(7): Left to right, instantaneous forecast error $T^{real} - T^{forecast}$ increases from $-3C$ to $3C$; Column (8): Marginal effect of negative forecast error; Column (9): Marginal effect of positive forecast error; Top to bottom: Functional forms of Poly 3, Bins and Cubic Splines; $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

Figure 2.B.3: Interactive Regression with Sample Restrictions



(a) Trim Temperature Outliers

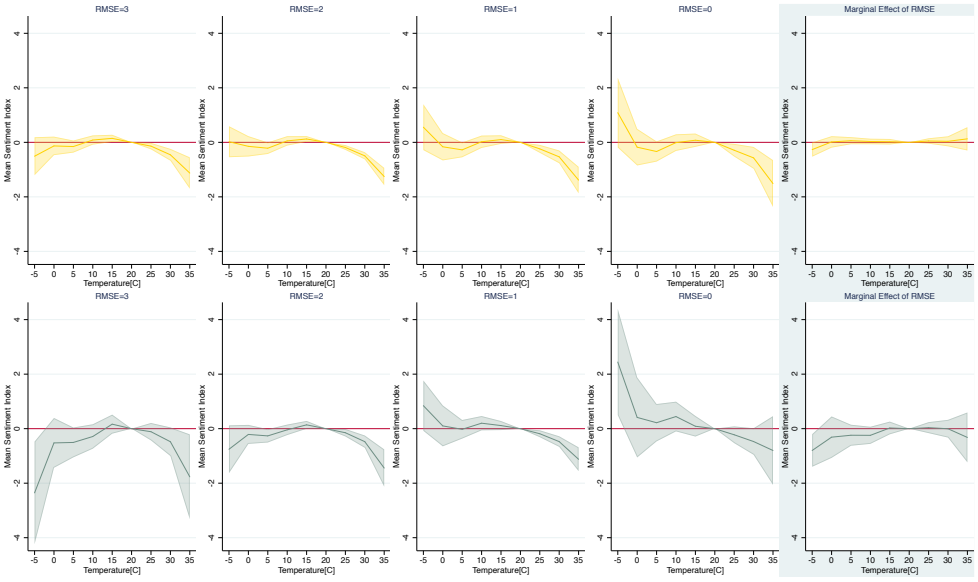


(b) Trim Forecast Error Outliers

Note: Column (1)-(7): Left to right, instantaneous forecast error $T^{real} - T^{forecast}$ increases from $-3C$ to $3C$; Column (8): Marginal effect of negative forecast error; Column (9): Marginal effect of positive forecast error; Top to Bottom: Covariate $|T^{real} - T^{forecast}|$ non-rationalized (top) and rationalized (bottom); Trimming either temperatures or forecast errors at top and bottom 1% each; $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

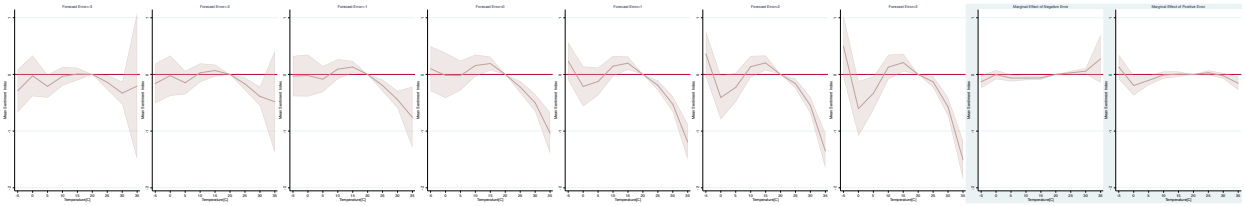
2.C Other Explorations

Figure 2.C.1: Interactive Regression with Long Run Forecast RMSE



Note: Column (1)-(4): Left to right, long run forecast RMSE $\sqrt{\frac{1}{N}\sum_t T_t^{real} - T_t^{forecast}}$ decreases from $3C$ to $0C$; Column (5): Marginal effect of RMSE; Top to bottom: Covariate RMSE with non-rationalized (top) and rationalized (bottom) forecasts; $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

Figure 2.C.2: Interactive Regression with Naive Forecast Error



Note: Column (1)-(7): Left to right, $T^{real} - T^{forecast}$ increases from $-3C$ to $3C$; Column (8): Marginal effect of negative forecast error; Column (9): Marginal effect of positive forecast error; Covariate $|T^{real} - T^{forecast}|$ with naive forecast based on AR(7) prediction; $T = Tavg$, reference bin $[15C, 20C)$; The label of T on horizontal axis indicates the temperature bin $[T-5, T)$, with the starting bin $(-\infty, T)$ and ending bin $[T-5, \infty)$; 95% confidence interval is shaded.

Table 2.C.1: Precipitation Category Response

	(1)	(2)	(3)	(4)	(5)	(6)
$BadWeather$	-0.187***	-0.188***	-0.305***	-0.298**	-0.307***	-0.336***
	(0.024)	(0.024)	(0.112)	(0.146)	(0.047)	(0.029)
$BadWeather \times Error$			0.393	0.335	0.324***	0.247***
			(0.378)	(0.444)	(0.110)	(0.021)
Control	No	Yes	Yes	Yes	Yes	Yes
Covariate	NA	NA	City-Year (2011)	City-Year (2014)	City-Month (2014)	City-Day (2014)
N	39529	39241	39241	39241	39241	39241
Adj. R^2	0.654	0.654	0.654	0.654	0.654	0.656

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.C.2: Precipitation Category Response, Capital Cities Only

	(1)	(2)	(3)	(4)	(5)	(6)
$BadWeather$	-0.194***	-0.194***	-0.482**	-0.771**	-0.327***	-0.331***
	(0.041)	(0.041)	(0.219)	(0.330)	(0.084)	(0.052)
$BadWeather \times Error$			0.968	1.789	0.365*	0.248***
			(0.781)	(1.061)	(0.198)	(0.036)
Control	No	Yes	Yes	Yes	Yes	Yes
Covariate	NA	NA	City-Year (2011)	City-Year (2014)	City-Month (2014)	City-Day (2014)
N	39529	39241	39241	39241	39241	39241
Adj. R^2	0.654	0.654	0.654	0.654	0.654	0.656

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Chapter 3

The Effect of the End-Number License Plate Driving Restriction on Reducing Air Pollution in China

Abstract

This project explores the long-run efficiency of the end-number license plate driving restriction in China, a traffic control policy aiming at reducing city level air pollution. Difference-in-difference (DID) regression analysis is conducted for a panel including 9 cities implementing this long run policy staggeringly over the period 2008-2013, contrasting to the set of control cities in China without the continuous implementations of this policy over the years 2005-2015. Consistent with previous studies on road rationing, this project has shown that the driving restriction policy on average is not improving air pollution in China in the long term. Quantitatively, the policy would reduce the daily city level AQI (air quality index) by a statistically insignificant 0.04% at most, translating to welfare gains of only 0.08 dollars per person per year and 0.004 life years per capita.

3.1 Introduction

Environmental issues, especially air pollution, have become quite serious concerns of the general public over the past few decades in China. When the country is developing with a fast pace of economic growth, worries around the proven health and mortality risk related to air pollution continues to grow. Correspondingly, the central and local governments have become increasingly progressive in issuing various environmental policies in order to render such concerns effectively. Among those policies, since car emission has long been suspected

to be one of the major sources of air pollution in city areas, road rationing, particularly in the form of the end-number license plate driving restriction, has become one of the most popular choices for local policy makers. This policy mainly limits the number of private owned vehicles driving on roads in city areas by forbidding sets by their end-number of plates, usually announcing its target to reduce air pollution and traffic congestion simultaneously. It has been one of the environmental policies affecting general public implemented earliest and longest since 2008, with its usage extending to many major cities across the country over the years. Due to its nature directly affecting potentially large population with private vehicles of any city implementing the policy, the efficiency of this kind of road-rationing policy has long been debated over, from whether there are any significant and lasting effects of air pollution reduction, to whether these policies sustain through the high costs of constraining private transportation, vehicle sales and other normal form of economic activities.

Road rationing has been implemented all around the world, both in developed and developing countries. Short run road rationing with large proportion (usually 50%) of private vehicles baned from road has frequently been noted as effective, thus often applied (Chen etal. 2013).[1] Long-run version of the policy banning less cars on major roads (usually 20% on working days) for continuous time span over months and years is less adopted on a large scale and shows more ambiguous effects on long term air pollution. One of the most known study about long run road rationing has been set in Mexico city where the policy is shown to have little to no immediate effects though the policy implementation being pretty strict

[1]Additional news article: https://www.nytimes.com/2014/03/18/world/europe/fighting-pollution-paris-imposes-partial-driving-ban.html, Paris partial driving ban 2014.

(Davis, 2008). The near-zero results are robust on multiple pollutants using regression discontinuity (RD) designs, verifying the ineffectiveness of the policy in immediate short-run air pollution reduction. Similar RD designs have been widely adapted in other studies of this kind of policy around the world, but reach heterogeneous conclusions regarding its efficiency in different countries. In Quito, capital of Ecuador, a significant medium-size effect reducing 9%-11% CO level has been found (Carrillo, Malik and Yoo, 2016).

Within China, existing studies have mainly been focusing on the capital city of Beijing where the policy is first implemented with strict guidelines, affecting the most population among similar policy cities, and have attracted most public and media attentions throughout the years. Long run continuous end-number license plate driving restriction of Beijing is noted to reduce up to 21% air pollution from 2007 to 2009 (Viard and Fu, 2015), but other literature questions the RD design and propose the actual reduction being much closer to zero (Sun, Zheng and Wang, 2014). Not as many literatures occur for other policy cities in China besides Beijing. Among those existing literatures, a study of Lanzhou in Middle-West China has found some short run but little long run effects on city air pollution reduction (Huang, Fu and Qi, 2016). There are also some evidence of pollution reduction is found in Hangzhou, East China (Ye and Zhuo, 2018). However, a reverse effect of the policy increasing air pollution instead has been noted in Chengdu, South-West China from 2011 to 2013 (Xu and Hou, 2015).

In this project, I will analyze the average long term lasting effect of the long run driving restriction across all policy cities in China using a difference-in-difference (DID) design. The

results would indicate whether this driving restriction has been effective on improving air quality over various cities across China, and from which I will discuss briefly what are the cost and trade-off for this policy. Since the long run end-number license plate driving restriction has been implemented gradually over a long time frame of a decade, I would focus on the long run effect of this policy which is more likely to create greater welfare impacts for the general public. Overall, my results show that the policy has little effects on average in reducing air pollution in China, translating to only small gains in currency values and public health, and thus likely unable to cover the potential high cost of implementations and negative shocks to transportations and auto markets.[2]

This project would differ from previous literatures through analyzing the average effect of the long run road rationing strategy across different cities of China, while existing studies either focus on short run odd-even rationing during special events like the Olympic Game of 2008, when more than one policy interventions were happening simultaneously (Chen etal, 2013), or only study the short run effect of the policy in a single city with time series analysis. There is (so far in my search) not enough evidence of panel studies on the average effects of this policy in different cities of China comparing with those not adapting the policy, for which my analysis will fill in the gap. I switch focus from the widely adapted regression discontinuity design in time (RDiT) to a panel design difference-in-difference, which would allow cross-sectional comparisons between policy and control cities. Unlike RD, this design

[2]This project is built on my 2014 undergraduate thesis paper analyzing the immediate direct effect of end-number license plate driving restrictions on air quality in seven Chinese cities, applying regression discontinuity design over a maximum time span of two-year window centered around the time of the policy change. Heterogeneous effects have been found for different cities, revealing the efficiency of this specific policy may have been ambiguous. This set of RD results are also replicated in **Section 5**.

is also less reliable on the policy implementation dates where they are not always carried out sharply as announced by the local governments, but rather gradually from days before to days after.

This paper proceeds in the following order. **Section 2** will introduce policy background and data. **Section 3** will consider empirical strategies. **Section 4** will enlist results of the main specifications. **Section 5** will explore alternative specifications including the RDiT design. **Section 6** will approach welfare estimates and extensions of the policy effects in other aspects besides air pollution reduction. **Section 7** will conduct the robustness checks with alternative air quality measures. **Section 8** will conclude with discussions and future remarks.

3.2 Data

3.2.1 Policy Background

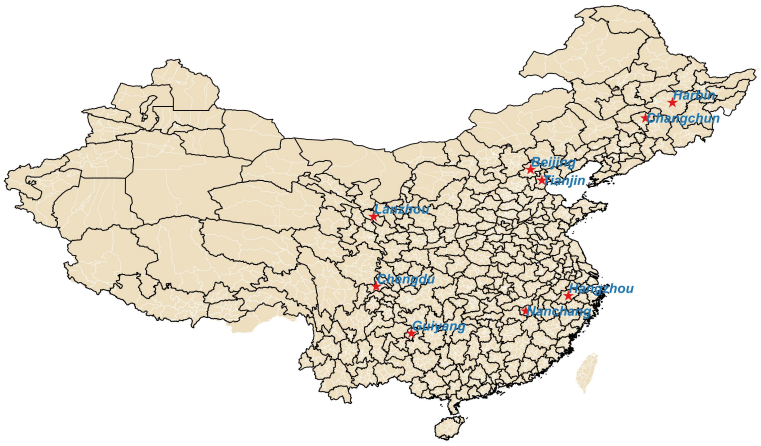
For this project, I focus on the continuous long-run version of the end-number license plate driving restriction policy. It is a design that is not as strict as the odd-even rationing often applied when there is city alert for high level of air pollution, banning a smaller proportion of private cars on road, but with no explicit termination dates into the future. In China, this almost always means that the policy is in effect once it has been announced. Details of the policy are refined slightly once per year, but among all the policy cities I look at it has not been terminated until today. From local city governments' official documents and

authorized news reports, I find summary of the long-run end-number license plate driving restriction in 9 cities of China before 2015, as showed in **Figure 3.1**:

1. Beijing (since October 11, 2008), capital city of China, a municipality city.
2. Nanchang (since June 22, 2009), capital city of Jiangxi Province.
3. Haerbin (since April 10, 2010), capital city of Heilongjiang Province.
4. Changchun (since May 4, 2010), capital city of Jilin Province.
5. Guiyang (since October 1, 2011), capital city of Guizhou Province.
6. Hangzhou (since October 8, 2011), capital city of Zhejiang Province.
7. Chengdu (since April 26, 2012), capital city of Sichuan Province.
8. Lanzhou (since June 1, 2013), capital city of Gansu Province.
9. Tianjin (since December 16, 2013), a municipality city.

This selected set of 9 cities are from across the country, but excluding those temporarily adapting road rationing during international events (e.g., Guangzhou and Jinan), or those only put restrictions on non-local plates (e.g., Shanghai), or those only using end-number plate control on bridges linking to other cities (e.g., Wuhan). These 9 policy cities have been adapting road rationing on major roads of their city centers, banning one-fifth of private cars from running during rush hours each workday between 7am and 8pm, excluding weekends and holidays. This make 20% cars in ban, and the set of bans rotates around the week by their end-numbers of license plates. For example, number 1 and 6 cannot drive on road Monday, then number 2 and 7 cannot drive on road Tuesday, and so on. Public transportations, such as buses and taxis are exempted from the restriction (except for the city of Lanzhou

Figure 3.1: Map of Policy Cities



Source: State Bureau of Surveying and Mapping

which also imposes restriction on taxis). By the definition of private cars, civil use vehicles that are not owned by individuals (e.g., school buses or corporate use vehicles) may register with local transportation department so they will not be restricted by the policy. Most of these policies are set by the local government and announced months in advance. Violators of the driving restriction once being caught are fined about 100 Yuan the first time with negative record added on driver's history report that may lead to cancellation of drivers' licenses, and fine increases if caught a second time in the same day. Violators can be caught by either local traffic police or traffic cameras around the cities.

3.2.2 Air Quality Measure

The main city level air pollution measure I choose to use in this paper is AQI, short for Air Quality Index. It has been the main index known and published online in China by Ministry of Ecology and Environment. AQI is an aggregate measure of air quality, considering multiple pollutants at once. Before 2013 the Ministry adapt Ambient Air Quality Standards GB 3095-1996 where pollutants in consideration are SO_2, NO_2, $PM10$. After 2012 the standard changed to GB 3095-2012 gradually [3], when pollutants considered for AQI measure further include $PM2.5$, O_3 and CO[4].

The formula for AQI computation follows linear extrapolation from individual pollutant concentration to the scale of AQI, and then taking the maximum over all pollutants in measurement[5]:

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_P - BP_{Lo}) + IAQI_{Lo}$$

$$AQI = \max\{IAQI_1, ..., IAQI_n\}$$

Here $IAQI$ is individual pollutant P's AQI index, Hi and Lo identifies the range pollutant concentration C_P in unit μ_g/m^3 falls between in the concentration thresholds BP from the conversion standard table. For both old and new standards of AQI, see tables **Figure**

[3]The goal is to alter all monitoring station data to the new standard by 2016, but major cities like Beijing start shifting to new standard since 2013.

[4]Source: http://www.bjepb.gov.cn/bjhrb/xxgk/fgwj/qtwj/zcjd/608633/index.html, http://kjs.mee.gov.cn/hjbhbz/bzwb/dqhjbh/dqhjzlbz/201203/t20120302_224165.shtml.

[5]Souce: Technical Regulation on Ambient Air Quality Index HJ633-2012, published by Ministry of Ecology and Environment of the People's Republic of China.

3.A.1 and **Figure 3.A.2** in **Appendix**. After converting all concentrations to $IAQI$ values for each of the n individual pollutants, max of the $IAQI$ is taken as the overall AQI index. The pollutant with maximum IAQI is called the major pollutant.

I process a web-scraped dataset [6] from the open source of Ministry of Ecology and Environment of the People's Republic of China. I drop 3 observations with negative AQI which shall be bounded by 0. The Ministry publishes real time daily average AQI with an indicator of major pollutant, from January 2000 to February 2015. When multiple monitoring stations are recorded in one city, the average AQI across all stations covered is recorded as the city measure. Number of cities in this dataset grows from 43 to 367 over the years as more monitoring stations are being built, covering most prefecture level cities and then some county level cities in China. I restrict my sample range to start from year 2005, so that the number of prefecture level cities (all the treatment cities are at prefecture level) recorded per year is 84 or above, covering the whole country.

Due to the major change of AQI standard calculation by the Ambient Air Quality Standards (GB 3095-2012 in place of GB 3095-1996) end of 2012, I only take the AQI from the Ministry website (before 2013 it has the alternative name of API, Air Pollution Index) from 2005-2012. For the rest of the years I cover 2013-2015, I obtain the by-station by-pollutant type daily concentration air quality records also scraped from web and shared by Professor Guojun He of Hong Kong University of Science and Technology. Station level daily AQI is

[6] Shared by an online coder with the agreement of usage, https://www.gracecode.com/aqi.html, author email lucky@gracecode.com.

directly computed from the concentration of three individual pollutants using the pre-2013 standard (**Figure 3.A.1** to keep consistency of dataset. Then I overlap the 1737 monitoring station locations by latitude and longitude with shapefile of prefecture level cities in China, average across all stations within boundaries of cities identified by QGIS to compute the city level daily average AQI. Identifier of major pollutant is approximated by recording the pollutant that become of the major one in the most stations within the city boundaries.

3.2.3 Robustness, Covariates and Extensions Data

Due to somewhat questioned credibility of published air quality data in China when local monitoring stations may manipulate and under-report true air pollution level so the city government would have air quality seemingly meeting with national standard, alternative data sources will be accessed for robustness checks in **Section 7**. Those include hourly *PM2.5* concentration measures from US Embassy in 5 Chinese cities[7], the NASA Anthropogenic Aerosol Optical Depth (AOD)[8], and annual average daily *PM10* concentration by city data from Professor Michael Greenstone. For more descriptions on these datasets, see **Section 7**.

For control covariates, I required permissions to use spatially aggregated population weighted climate data from the EPIC Climate Impact Lab, with which I generate daily ERAI temperature and precipitation at prefecture city level. GIS shapefile of the country up to ADM3 (county/district level, represented by 6-digit district code) is obtained from National Catalogue Service for Geographic Information, published by State Bureau of Sur-

[7]Souce: Kaggles.com
[8]Source: https://neo.sci.gsfc.nasa.gov/view.php?datasetId=MODAL2_M_AER_OD

veying and Mapping in 2017 (the latest version) and boundaries effective to around 2015[9]. The centroid latitudes, longitudes and areas of prefecture cities in China are then computed in QGIS from the shapefile at ADM2 level, merged with the first 4-digit prefecture level city code 2015-2018 published by the Ministry of Civil Affairs[10], projection system WGS84.

For extensions analysis on the policy effects on aspects other than air pollutions, I obtain annual city level traffic, transportation, demographic and macroeconomic variables from electronic tables published by China City Statistical Yearbook, and province level annual auto market data from China Auto Market Almanac[11]. I also obtained labor time-use survey data from China Health and Nutrition Survey (CHNS)[12], with ADM2 level location crosswalk compiled from the appendix table p659-p662 of Ge (1998).

3.3 Empirical Strategies

3.3.1 Event Study

First analysis I conduct will be an event study. The event study scheme aims to study event window effects of the outcome AQI variable, as an indication of whether pre-policy effects take place early before the official execution dates and whether post-policy effects exists and last:

[9]Website: http://www.webmap.cn/commres.do?method=result100W
[10]Website: http://www.mca.gov.cn/
[11]Source: http://tongji.cnki.net
[12]Source: https://www.cpc.unc.edu/projects/china

$$AQI_{it} = \alpha + \sum_{k=\underline{k}}^{\bar{k}} \beta_k D_{it}^k + \mathbf{X}_{it}'\gamma + \delta_t + \phi_i + \epsilon_{it}$$

For which $D_{it}^k = \mathbb{1}_{t=c+k}$ where c is the first day of the policy being implemented at city i. Here I will take a window of $[\underline{k}, \bar{k}] = [-30, 30]$ for 60 days around the policy implementation date for each city, and the key coefficients β_k determines the average effect of event day k within this window comparing with control cities without policy implementations. I bin up and down the first and last bin such that $D_{it}^{\bar{k}} = \mathbf{1}_{t \geq c+\bar{k}}$ and $D_{it}^{\underline{k}} = \mathbf{1}_{t \leq c+\underline{k}}$. For the purpose of normalization, I omit the dummy for first policy date c, i.e., D_{it}^0 is omitted and serve as the reference date.

Since my study will focus on the long term effect of the driving restriction, I will also conduct a longer event window analysis extending to 30 months (two and a half years) before and after the policy dates, where D_{it} is now monthly instead of daily dummies. Under the monthly setting, policy date is redefined as the first month with at least half a month treated under the driving restriction.

The covariate control vector $\mathbf{X}$ includes climate variables (daily or monthly average temperature and total precipitation), indicators for major pollutants being SO_2, NO_2 or $PM10$. Day and city fixed effects are added. I also control for city-by-year fixed effects, capturing macro city level annual shocks such as GDP growth, policy shocks, population changes, etc.. Standard errors are clustered at city-by-year level, allowing correlations within same city of each calendar year.

3.3.2 Difference in Difference (DID)

The main identification strategy this project use will be DID, which would estimate the average effect for the end-number license plate control policy over all 9 treatment cities in China comparing with control city sets. The following regression will be run on full set of treatment and selected control cities:

$$AQI_{it} = \alpha + \beta Policy_{it} + \mathbf{X}'_{it}\gamma + \delta_t + \phi_i + \epsilon_{it}$$

Where the key policy indicator $Policy_{it}$ equals to 1 when the road rationing policy of city i at time t is in effect. Control vector $\mathbf{X}$ again includes climate variables (daily average temperature and daily total precipitation), indicators for main pollutants being SO_2, NO_2 or $PM10$. Day and city fixed effects are added to maintain the DID design, and I again control for city-by-year FE for other city-year level economics or demographic macro shocks. Standard errors are again clustered at city-by-year level. The key assumption of DID, namely the parallel trend pre-treatment, can be tested through event study of previous subsection if pre-policy trend levels around the zero axis.

Besides this main specification, I will also replicate the usual RDiT (regression discontinuity in time) for immediate direct policy effects in shorter terms for individual policy cities, and use extensions of synthetic controls to estimate the long run policy effects in different policy cities. Details on those alternative specifications are in **Section 5**.

3.3.3 Control Cities Selection

For both event study and DID, a comparable set of control cities need to be selected in contrast to the 9 treatment cities. Before the selection process, I first keep my control cities selection only among those with at least one day of non-missing AQI measurement per year covering all 11 years 2005-2015. That counts to a pool of 75 control cities. To make sure any results in the next sections will be robust across selections, I will adapt four different sets of control cities:

1. Only include the set of 9 treatment cities which adapt the long-run driving restriction, based on the argument that treatment cities are comparable to one another but not to cities without such policy.

2. Select control cities within the same provinces as the treatment cities[13]. Same-province cities are grouped together as controls because they may usually share common regional time-varying policy and macroeconomic shocks.

3. Take 5 nearest matched neighbors of each treatment city, matching on covariates that likely define the economic conditions, the public and government emphasis on environmental issues, and geographical distances to the treatment city. The covariates I use for matching include 2005-2007 prior policy average demographic, economic and geographic variables, which are population, population density, primary, secondary and

[13]For Beijing and Tianjin which are municipalities directly under the central government, the province they geographically have been enclosed by is Hebei. Also many economic and environmental policies are implemented to the whole area of Beijing, Tianjin and Hebei, which is called the the Beijing-Tianjin-Hebei region.

tertiary industrial share of GDP of the city, GDP per capita converted to 2015 value, unemployment rate, private and self employment rate, ratio of labor force to city population, ratio of number of cell phone registers over city population, ratio of number of Internet registers over city population, road area per capita, and then the spatial distance between centroids of cities, and the indicator on whether the candidate control city is within the same province of the treatment city. The metric used for determining nearest neighbor is the Mahalanobis distance. Each control cities matched as 5 nearest neighbors takes a weight of 0.2 per matching to a treatment city.[14] For alternative matching to different number of nearest neighbors, and with different set of covariates, see **Section 5** for checks.

4. Include the full sample of all 75 control cities available all time span 2005-2015 in my air pollution dataset.

In later sections, most regression analysis will be conducted on all four sets of control cities. Note that the set of control cities, however, may not be entirely unaffected by some sort of driving restriction. Many of them occasionally use short run end-number license plate control during major events of the cities, or on non-local vehicles, or temporarily during days and months where air quality is bad. This would be more common among cities within the same province of the treatment cities, because the treatment cities are always capitals and policy makers in other cities tend to "follow" their policies. Overall, the number of cities ever implementing some sort of road rationing is growing over years in China, and the city-

[14]The design is similar to Cicala (2015), though I do not limit the selection pool of control cities to be within a certain distance to treatment cities because cities are much apart in my sample with large variety in their areas and shapes. Therefore, for me determining the cutoff distance is not quite intuitive as the reference paper. Instead I just incorporate the distance measure into my matching scheme.

time specific details of each regulations have grown more complicated[15]. However, high cost of implementation and potential opposition of the public has caused the number of cities adapting long-run continuous version of road rationing not grow as much. That is one of the reason why I restrict the time range of this study until 2015, before more of these noise enter the treatment and control assignments in most recent years.

But my empirical strategies will still rely on the assumption of average effect of short-run road rationing being on average similar in treatment and control cities. One may reasonably argue that treatment cities have more emphasis on air pollution controls, thus they may be applying short-run road rationing more frequently than control cities. This will make my estimate biased as a combined effect of both the long-run policy and more frequent adaptation of the short-run policy.

3.3.4 Unconditional Quantile Regression

Following a similar breakdown approach of the DID analysis to the US Clean Air Act in Currie, Voorheis and Walker (2019), I conduct the unconditional quantile regression (UQR) approach replacing LHS with 19 RIF (re-centered influence function) quantiles from 5% to 95%:

$$\text{RIF}(AQI, q_\tau) = q_\tau + \frac{\tau - \mathbf{1}_{p \leq q_\tau}}{f_{AQI}(q_\tau)}$$

Where q_τ is the τ^{th} quantile function , and f_{AQI} is the density function of AQI pollution

[15]Source: http://zhengzhou.auto.sohu.com/20140425/n398804182.shtml.

measure. This regression would allow a distributional analysis of the policy impact based on 5% to 95% quantiles of pollution, for all four choices of control city sets.[16] with This would help answer questions regarding whether policy impacts varies based on current pollution levels.

3.4 Main Results

3.4.1 Balance Checks

Before any regression analysis, a set of balance checks are performed for the 9 policy cities, divided through those with early execution (year 2010 or before) and late execution (year 2011 or after). The early-group then includes Beijing, Nanchang, Haerbin and Changchun, and the late-group includes Guiyang, Hangzhou, Chengdu, Lanzhou and Tianjin. Covariates to check are various demographic, geographic, climate and pollution variables average across 3 three-year windows 2005-2007, 2008-2010, 2011-2013. As shows in **Table 3.1**, none of the T-test between the early and late groups have been rejected at 10% level. Also regress some of these major covariates on policy implementation dates[17], results of **Table 3.2** shows no significant coefficients for any covariates, suggesting that there is not rejection of null hypothesis that the date of policy implementation is independent on observables. However, note all these tests are quite underpowered with $N = 9$. Therefore, the results still cannot rule out possibilities of covariates imbalance by the implementation strategies of different policy cities.

[16]The set of quantile regressions is executed in STATA using existing package, source https://www.stata.com/meeting/chicago19/slides/chicago19_Rios-Avila.pdf.

[17]I have to cut down number of regressors since I only have 9 observations.

Table 3.1: Balance Table Check for Early and Late Policy Cities

Variable	Early Group Average	Late Group Average	Difference	P-Value
City Area (km^2)	33597.770	14815.223	-18782.547	0.194
Average Over Period, 2005-2007				
Daily AQI	80.667	84.759	4.091	0.702
Cellphone/Pop.	0.715	0.737	0.022	0.924
GDP per Cap. (2015 Yuan)	43381.335	42489.963	-891.372	0.948
Internet/Pop.	0.197	0.140	-0.057	0.330
Labor/Pop.	0.360	0.287	-0.074	0.551
Pop. Density (pc/m^2)	382.756	465.703	82.947	0.606
Population (10000 pc)	850.106	676.919	-173.187	0.462
Daily Precip. (mm)	1.991	2.655	0.664	0.493
Prim. Ind. Shr. (%)	8.240	4.923	-3.317	0.266
Prop. Private Employed (%)	33.125	34.893	1.768	0.733
Road Area Per Cap. (m^2)	8.178	9.298	1.120	0.426
Second. Ind. Shr. (%)	41.716	48.713	6.997	0.262
Daily Avg. Temp. ($^\circ C$)	10.054	13.435	3.381	0.343
Ter. Ind. Shr (%)	50.043	46.363	-3.679	0.594
Unemployment Rate (%)	3.176	2.980	-0.196	0.785
Average Over Period, 2008-2010				
Daily AQI	74.837	79.420	4.583	0.567
Cellphone/Pop.	0.969	1.095	0.126	0.574
GDP per Cap. (2015 Yuan)	53925.235	53095.286	-829.949	0.957
Internet/Pop.	0.215	0.201	-0.014	0.876
Labor/Pop.	0.367	0.366	-0.001	0.993
Pop. Density (pc/m^2)	397.240	477.939	80.699	0.632
Population (10000 pc)	878.266	695.663	-182.603	0.461
Daily Precip. (mm)	2.171	2.695	0.524	0.608
Prim. Ind. Shr. (%)	6.833	3.998	-2.835	0.252
Prop. Private Employed (%)	36.770	42.869	6.099	0.189
Road Area Per Cap. (m^2)	8.993	9.757	0.763	0.682
Second. Ind. Shr. (%)	41.851	47.758	5.907	0.396
Daily Avg. Temp. ($^\circ C$)	9.514	12.980	3.466	0.333
Ter. Ind. Shr. (%)	51.317	48.244	-3.073	0.688
Unemployment Rate (%)	3.624	2.407	-1.217	0.253

Table 3.1, continued

Variable	Early Group Average	Late Group Average	Difference	P-Value
Average Over Period, 2011-2013				
Daily AQI	73.622	77.442	3.820	0.611
Cellphone/Pop.	1.377	1.605	0.227	0.519
GDP per Cap. (2015 Yuan)	69597.887	71755.298	2157.410	0.916
Internet/Pop.	0.259	0.232	-0.027	0.726
Labor/Pop.	0.455	0.465	0.011	0.945
Pop. Density (pc/m^2)	403.762	493.875	90.113	0.602
Population (10000 pc)	889.008	717.473	-171.535	0.496
Daily Precip. (mm)	2.170	2.633	0.463	0.580
Prim. Ind. Shr. (%)	5.961	3.203	-2.757	0.207
Prop. Private Employed (%)	42.661	40.359	-2.303	0.583
Road Area Per Cap. (m^2)	12.231	11.203	-1.028	0.691
Second. Ind. Shr. (%)	42.102	46.481	4.379	0.556
Daily Avg. Temp. ($^\circ C$)	9.409	12.829	3.420	0.353
Ter. Ind. Shr. (%)	51.938	50.317	-1.622	0.841
Unemployment Rate (%)	2.533	2.063	-0.470	0.599

Note: Difference equals late group average minus early group average; P-value indicates T-test p-values between the group means; Annual city-level data source China City Statistical Yearbook; Daily AQI from Ministry of Ecology and Environment; ERAI climate data from Climate Impact Lab provides daily temperature and precipitation, aggregated from grid to prefecture level cities population weighted; GDP per capita in Yuan adjusted to 2015 by GDP deflator source World Bank.

Comparing treatment and control cities possess more power but also create more unbalance. In **Appendix Table 3.B.1** covariates balance checks are performed between treatment cities and the three sets of control cities without policy interventions for 2005-2007 average. The three sets of controls namely are 1) the within-province control cities; 2) the five-nearest-neighbor matched control cities and 3) full set of all control cities. From this table, we can see that in all cases there exist some socio-economical variables with mean statistically different between treatment and control groups, especially for daily average AQI when treatment cities always start with higher levels before 2007. That reveals partially the reasonings for treatment cities to adapt the policy in the first place, since they are in need of controlling for air pollution levels. However overall, we still see over half of the balance

Table 3.2: Balance Regression Check

	(1)
	Driving Restriction Implementation Date
Population Density (pc/m^2)	0.167
	(2.004)
GDP per Capita (2015 Yuan)	-0.003
	(0.031)
Unemployment Rate (%)	961.427
	(862.655)
Daily Average Temp. (C)	93.881
	(223.455)
Daily Precipitation (mm)	115.985
	(809.212)
Daily AQI	70.911
	(54.911)
Constant	8566.195
	(8704.082)
N	9
Adj. R^2	-0.808

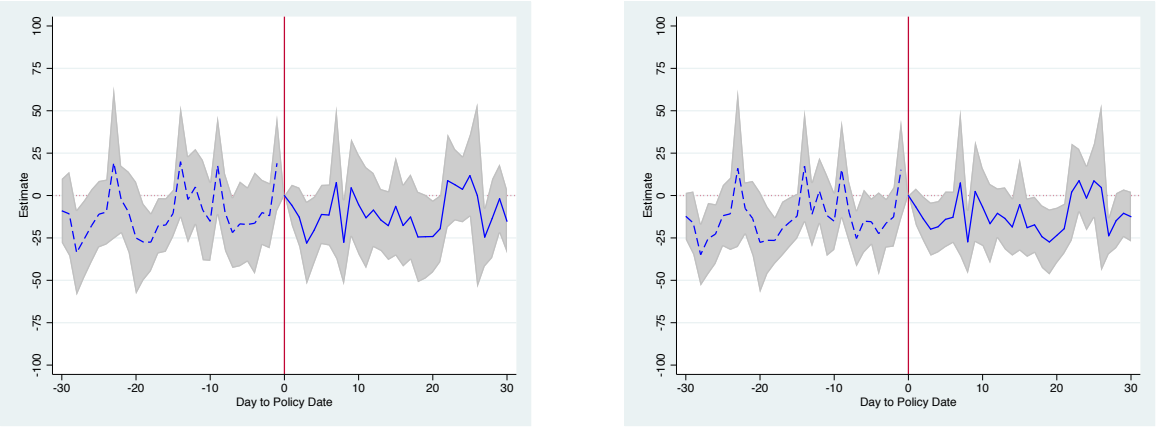
Note: Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressors are average city level across pre-period 2005-2007.

null test not being rejected. Also most of these unbalanced covariates, such as population size, ratio using Internet, ratio being in labor force, industrial structure, are likely to be more time invariants, thus they can be controlled by the city-by-year FE in later regression analysis.

3.4.2 Event Study

Figure 3.2 presents the result of event study with 60 days window, full set control cities, plotting coefficients $\hat{\beta}_k$ against k. The same event study plots with the other three options of controls are very similar under **Appendix Figure 3.C.1, 3.C.2, 3.C.3**. From the event study plots we can see several things. First of all, pre-policy trend almost never statistically significantly diverge from zero and is fluctuating along a flat level, therefore consistent with the parallel trend assumption for later DID analysis. Secondly, if we grade the level of the estimates before and after the policy change over the 60 days window, there seems no significant level difference on average. There is also no observable evidence of an average immediate effect of the policy, the drop of β_k right at the policy date seems more likely a noisy return-to-mean behavior from unexplained high peak one day right before the policy. Thirdly, majority of $\hat{\beta}_k$ regardless of pre- or post-policy are not statistically from zero, though the point estimate range from about -25 to 25. Confidence intervals are slightly tighter when control covariates are added, but relative magnitudes of point estimates are not shifted much. Overall this event study result implies pretty noisy but on average zero effects over the event window around policy implementation dates, and that casts some doubts on whether the policy is efficient in air pollution reduction even in the short run.

Figure 3.2: Event Study with Window of 60 Days, Full Controls



(a) Without Covariates

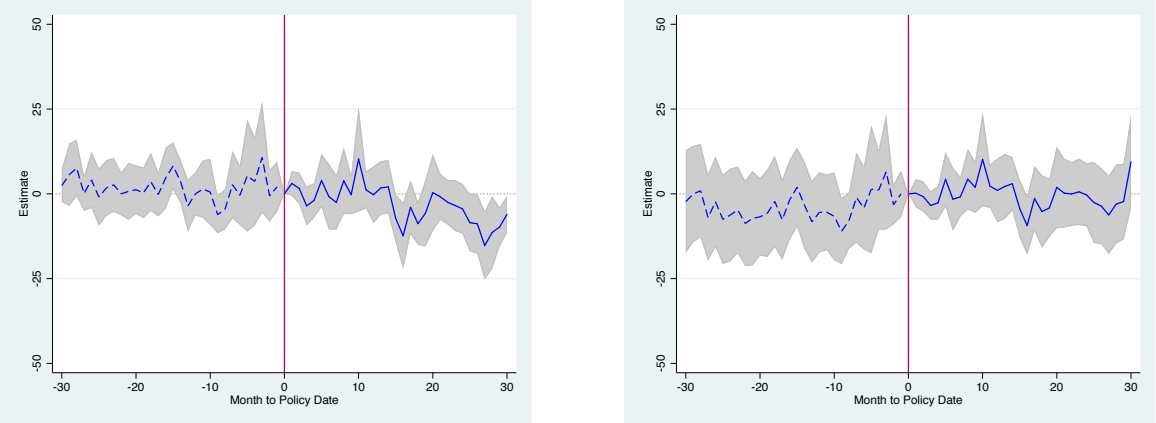
(b) With Covariates

Figure 3.3 shows the event study plots for 30 months window before and after the policy cutoff. In this longer range of time, we still see pre-policy zero trend consistent with parallel assumption. We also see more trend of decreasing AQI, thus improving air quality after about 1 year of the policy, but it is again quite noisy and fluctuating with 95% confidence intervals mostly covering the zero axis. In addition, the downwards trend disappears after covariates have been added. Therefore with those estimates, I can still claim that the event study in longer runs of time shows at least uncertain and small policy effect in air pollution reduction.

3.4.3 Difference-in-Difference (DID)

The main result of DID regression with the multiple options for control cities selections is presented in **Table 3.3**. This table shows on average there is no statistically significant decrease in AQI measure for cities implementing the long-run end-number license plate policy, comparing with those without the policy change, once controlling for full set of covariates.

Figure 3.3: Event Study with Window of 60 Months, Full Controls



(a) Without Covariates

(b) With Covariates

The no covariates version have some of the statistical significant negative estimates with considerable magnitudes when control cities include not only those with a policy treatment, but by balance checks in previous subsection those estimates are more likely biased without covariates. Once all covariates are added, all estimates across the panel are statistically insignificant and small in magnitudes regardless of the control groups selected. Even the 95% confidence interval lower bound never exceeds −5 across the panel.

Comparing with a panel data mean of 80.8 across the 9 policy cities before policy implementation, regressions with control covariates all show at most 0.04% decrease of AQI from point estimates, and maximum reductions of 4.4% to 5.7% by 95% confidence interval lower bounds. These numbers are confirmed with even lower estimates when replacing dependent variable with log AQI in **Table 3.4**. Point estimates on policy indicator become all positive after covariates added, and lower bounds of 95% CI ranges from only 3.2% to 3.7% AQI reduction. Thus overall, my DID main results show little to no average effects of this road rationing policy on air quality improvement in China.

3.4.4 Unconditional Quantile Regression

Unconditional quantile regression using RIF transformation has result presented in **Figure 3.4**, all four control city sets, with LHS quantiles on log AQI for the purpose of interpretation. In those four plots, blue bars are DID without covariates added, red bars are with covariates added. 95% CI are displayed as error bars.

UQR results shows the near-zero insignificant policy impact consistent across all quantiles for covariates added regressions, and for no covariates added estimates with only policy cities included. Negative policy effects shows up mostly only at high quantiles above 85% indicating long-run policy impact may be more significant over very high air pollutions, but with covariates in DID they are always statistically insignificant. Lower 95% CI bounds stays small negative as well across the control selections, mostly below 10% AQI reduction unless reaching high quantiles. Even for lower bounds without covariates, the magnitudes are generally not exceeding a 20% reduction unless top quantiles.

Table 3.3: Difference-In-Difference (DID) Regression Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Policy	0.259	1.217	-5.830***	0.068	-7.735***	-0.035	-6.066***	0.147
	(2.368)	(2.389)	(2.210)	(1.951)	(1.880)	(1.972)	(1.618)	(2.420)
	[-4.439,4.958]	[-3.524,5.959]	[-10.184,-1.476]	[-3.777,3.913]	[-11.431,-4.038]	[-3.912,3.841]	[-9.242,-2.890]	[-4.603,4.896]
Control Cities	Only Policy	Only Policy	Same Province	Same Province	Matched	Matched	Full	Full
Controls	No	Yes	No	Yes	No	Yes	No	Yes
FE	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date
N	35615	35615	81992	81992	136063	136063	324934	324934
Adj. R^2	0.285	0.380	0.317	0.450	0.326	0.449	0.341	0.467

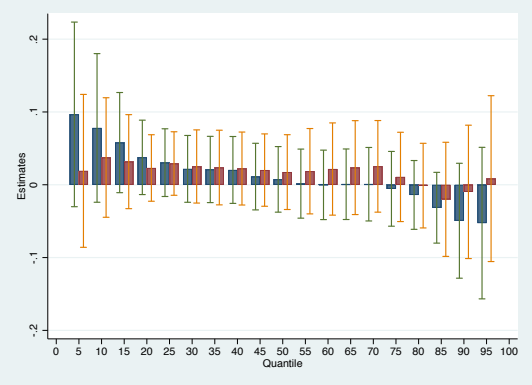
Note: Standard errors in parentheses, clustering on city-by-year level; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is daily average AQI; Controls include climate variables (daily average temperature and daily precipitation), city-by-year FE, and indicators for major pollutants.

Table 3.4: Difference-In-Difference (DID) Regression Results in Logs

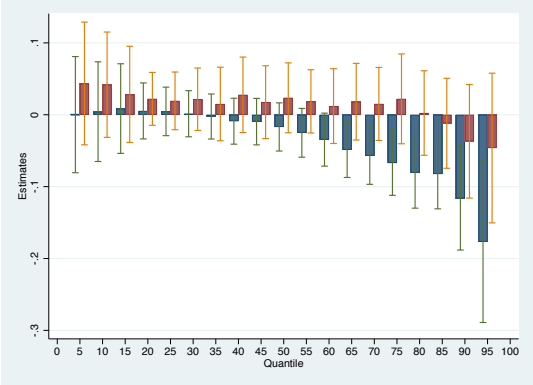
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Policy	0.014	0.015	-0.046**	0.008	-0.082***	0.003	-0.066***	0.003
	(0.029)	(0.023)	(0.023)	(0.023)	(0.021)	(0.021)	(0.019)	(0.021)
	[-0.043,0.071]	[-0.032,0.061]	[-0.092,-0.001]	[-0.037,0.053]	[-0.125,-0.040]	[-0.037,0.044]	[-0.103,-0.030]	[-0.037,0.043]
Control Cities	Only Policy	Only Policy	Same Province	Same Province	Matched	Matched	Full	Full
Controls	No	Yes	No	Yes	No	Yes	No	Yes
FE	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date
N	35615	35615	81992	81992	136061	136061	324932	324932
Adj. R^2	0.334	0.546	0.363	0.590	0.371	0.599	0.406	0.627

Note: Standard errors in parentheses, clustering on city-by-year level; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is daily average AQI; Controls include climate variables (daily average temperature and daily precipitation), and indicators for major pollutants.

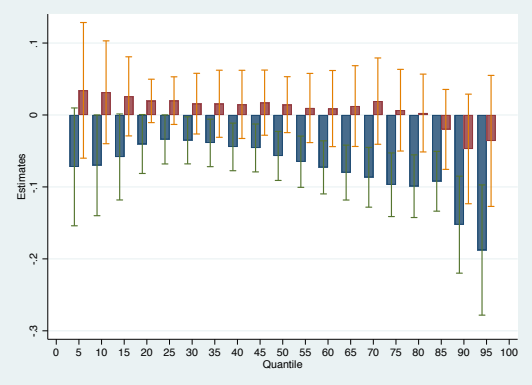
Figure 3.4: Unconditional Quantile Regression in Logs, 5%-95% RIF Quantiles



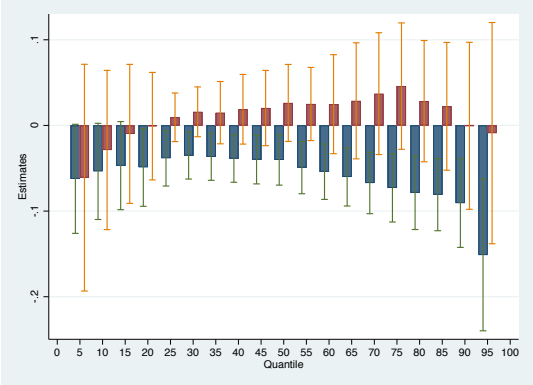
(a) Only Policy



(b) Same Province



(c) Matched



(d) Full

3.5 Alternative Specifications

In this subsection, I will conduct some exercises on alternative specifications that may be useful in supplementation of the results I achieve from main DID previous section. These alternative specifications will, besides reinforce the robustness of the main results, perform some tentative explorations on whether the small policy impacts achieved **Section 4** is a result of averaging heterogeneous significant positive and negative effects in different policy cities, or the small estimate generally apply to all policy cities alike.

3.5.1 Different Matching Controls

In my main **Table 3.3**, I choose nearest neighbor matching of 5 cities closest to each treatment city. To see weather the arbitrary choice of 5 nearest neighbor affect the main results, I also complete this exercise following Cicala (2015) varying the number of matching size to $m = 1, 5, 10$, and then rerun the main DID. Results in **Table 3.D.1** shows that, matching DID is not very sensitive to the matching sample size m, where point estimates on the policy dummy is always small in magnitude and statistically insignificant once covariates are added to the regressions. Also, the five-nearest-neighbor matching presented in main table is actually with the most conservative, greatest size negative point estimate and lower confidence interval bound across different m. Therefore, the matching results are stable, small and insignificant around zero.

In light of the balance checks, where matched controls and treatment groups are not completely balanced because I put more emphasize on matching with cities that are geographically closer to the treatment cities by including distance measures, I also conduct alternative matching without including the distance to the treatment cities or the same-province dummy indicator, such that we do not mechanically get control cities that are close to the treatment cities but instead those more similar in terms of socioeconomic factors. Results with different matching size $m = 1, 5, 10$ are presented under **Table 3.D.2**. While no-covariates regressions estimates become slightly smaller in magnitudes, estimates for with-covariates regressions increase in magnitudes and flip signs from main matching table. However, small insignificant estimates and similar size 95% CI lower bounds still persist

across all matching sample sizes after covariates are controlled. The main matching results are verified as stable.

Overall, the stable estimations lend some confidence to the matching algorithm I have performed. When in main result table matching gives the greatest negative estimate and confidence interval lower bound, I will be able to use this as the conservative "most" efficient pollution reduction estimate in later discussions.

3.5.2 Regression Discontinuity Design

Based on the work by Hausman and Rapson (2018), Regression Discontinuity in Time (RDiT) can help to break down and estimate heterogeneous policy impacts for different cities in the short run, immediately after the policy change. For a general RD design, the regression is run for each individual treatment city i:

$$\log(AQI_{it}) = \alpha_i + \beta_i Policy_{it} + f(R_{it}) + \epsilon_{it}$$

Where the key policy variable $Policy_{it}$ again equals to 1 when the policy of city i at time t is in effect. No control vector is included. Standard errors will take AR(1) to handle first-order serial correlations in time due to the nature of air pollution. f is a polynomial of the running variable, defined as the current date t minus the policy cut-off date c. The polynomial order of this RD running variable is chosen to be 1 (linear local regression) for main analysis, asymmetric by left- and right- of the policy date. A uniform kernel is adapted

weighting observations within the bandwidth equally. For choice of bandwidth h where the RD is designed upon interval $[c-h, c+h]$, I take trials of $h = 30, 180, 360$ days, from one month to two years around the policy implementation dates. Since separate cities have different historical AQI when the policy become in effect, log AQI is used as the dependent variable in this part.

Results are presented in **Table 3.E.1**. Results of the middle window $h = 180$ (a year around policy implementation dates) are also plotted under **Figure 3.E.1**. From those table and plots, we can see that the only city with a negative significant sizable impact of 30%-35% AQI reduction consistent for all bandwidth choices is Lanzhou, followed by Changchun, Chengdu, Harbin and Nanchang with significant and large estimate (reaches about 30% AQI reduction) for some of the bandwidth choices. Beijing, Tianjin and Hangzhou seems to have mostly counter-projected immediate result of increasing AQI by considerable magnitudes at least 20%, while Guiyang has statistically insignificant estimates for the policy impacts throughout the bandwidths. If we look at the $h = 180$ window RDiT plots, among the cities showing negative significant immediate policy impacts on AQI reduction, Changchun, Nanchang and Lanzhou in addition features the trend with increasing AQI within 180 days after the policy, showing possibilities of attenuating the policy effect in the long run. Only Chengdu in this regression seems to be pertaining a steady lowered AQI level after the policy. However, note the shape of RD plots are only illustrative, as imposing higher polynomials may reveal very different results.

For robustness checks running variable local polynomials order up to 4 are also performed

for bandwidth of 180 in **Table 3.E.2**. As expected, when polynomial order increases, some of the significant RDD estimates attenuate across the board. Among the cities identified with negative instantaneous reduction of AQI after the policy, the city of Changchun, Harbin, Lanzhou and Nanchang maintain a negative sizable impact but sign flips, size shrinks and significance disappears for some of the polynomials at least. Chengdu, which has a negative significant reduction effect around 20% from local linear specification, has the effect dropped and even flipped to positive significant at higher polynomial orders. Beijing, Tianjin and Hangzhou still seems to have mostly counter-projected immediate result of increasing AQI while statistical significance also reduced with higher polynomial orders, while Guiyang has statistically insignificant estimates again throughout the orders, though the estimates are shown to be large and negative at higher polynomial orders.

Overall, the RDiT results have presented some evidence of the heterogeneous immediate effects of this driving restriction policy across the 9 different cities. However comparing with DID, we cannot completely rule out the possibility of regional time trend, and we do not know whether the short term immediate effects can imply what happen in the long run. Any negative effect on AQI may get reverted later on, and any positive or insignificant effects may gradually phasing into effectiveness. The potential fuzziness of policy implementation dates based on how strict the penalization on violation is enforced over time may also bias the RDiT estimates.

3.5.3 Synthetic Controls

I also conducted the synthetic control method on main DID design. In this exercise, synthetic controls are constructed separately for each of the 9 treatment cities, according to all pre-policy period AQI measures since 2005 until the policy date. For now the optimization weighting matrix is taken to be the identity matrix.

With the synthetic control weights obtained, event study is rerun pooling treatment cities and synthetic cities weights. Results are shown in **Figure 3.F.1** and **Figure 3.F.2**. These figures are virtually the same as my original event study plots in **Section 4**, implying small pre- and post- effects of the policy on AQI reductions. The same DID design with synthetic control weights are run for both pooled and individual cities using log AQI, for different cities have different bases. Results are presented in **Table 3.F.1**. For the pooled DID with all treatment and synthetic control cities, the average effect is again small, positive and insignificant with 95% CI lower bound sit at 3.9% reduction similar as in the main table **Table 3.4**.

But in addition, we can see heterogeneous long run policy effects across different cities, though four out of nine estimates remains statistically insignificant. Among all the policy cities, Beijing, Changchun, Guiyang and Hangzhou have negative policy impacts on AQI, with Guiyang and Hangzhou statistically significant and large in magnitudes featuring around 10% reduction of long term daily AQI. Meanwhile estimates of policy dummy on Chengdu, Harbin, Lanzhou, Nanchang and Tianjin show positive signs, with Chengdu,

Harbin and Tianjin statistically significant and large above 10%. The maximum AQI reduction size across all cities is 11% at point estimate (Hangzhou) and 19% at lower CI (Beijing). These reductions are not a super big effect but also not minor.

These results do not seem to be dependent on the initial AQI level before policy is implemented, since the higher AQI city Hangzhou and lower AQI city Guiyang has the similar size negative significant policy estimates, while Harbin with a similar historical AQI pre-policy as Hangzhou instead has a large significant reverse sign estimate. These estimates also show little trace being dependent on the strictness of policy implementations, since the two adjacent municipalities of Beijing and Tianjin likely with similar mechanisms of policy implementations become to have reverse sign estimates. Results are also distinctively different from the RDiT estimates for immediate policy impacts, showing some evidence of the road rationing policy being more effective or less so in the short comparing with long run. For example, Hangzhou is with a positive significant AQI increase over 20% by RDiT instead has a long run DID estimate of 11% AQI reduction.

Overall, these results with synthetic control depict the possibilities of the end-number license plate driving restriction not being ineffective generally across all cities, but instead is capable of leading to heterogeneous large negative or positive effects on city level AQI. However, the reasons why some cities have the same policy work towards the government goal of AQI reduction while others work completely in the reverse direction remains unknown by this section, and will be explored slightly over the next. Note the method of synthetic controls matches only on pre-policy AQI trends, which may or may not be more defensible

as AQI is an one-dimensional measure. The estimates from this method can be less reliable if one believe the constructed synthetic control cities are not actually comparable to the treatment cities.

3.6 Extensions

3.6.1 Conversion to Benefit Value

Welfare cost and benefit analysis of end-number license plate control policy involves conversion of the main estimates previous section to currency values, which is not quite straight forward since AQI is a compounding measure from concentrations of different pollutants and only the health risk and costs of individual air pollutants have been estimated in previous literatures. I therefore only conduct a rough correlation estimation, regressing the daily *PM10* annual average measure with the daily AQI data averaged to city-year level controlling for yearly aggregated daily average temperature and precipitation. The slope of *PM10* concentration on AQI is 1.800. Comparing with the scaling table in Appendix **Figure 3.A.1**, this estimate is sensible as AQI is usually scaled lower from *PM10* concentration in $\mu g/m^3$. Then I use the Ito and Zhang (2016) estimate of willingness-to-pay (WTP) reducing $1\mu g/m^3$ of *PM10* permanently by paying 1.34 USD per year per person, which by current exchange rate is about 9.26 Chinese Yuan. Taking the DID estimate under **Table 3.3** with full sets covariates, point estimates of average reduction on AQI worth $-2.72 - 0.08$ dollars and $-20.28 - 0.58$ Yuan. Considering the 95% CI lower bounds, the maximum reduction of AQI would worth $8.50 - 11.10$ dollars and $58.74 - 76.72$ Yuan per year per person. These

numbers are small and mostly affordable in Chinese households. In terms of other welfare measures, the summary of EPIC[18] claims that "sustained exposure to an additional $10\mu g/m3$ of *PM10* reduces life expectancy by 0.64 years". Thus the result of this paper maps to a per capita life expectancy gains of $-0.140 - 0.004$ years by point estimates, or maximum gains of $0.405 - 0.530$ life years by 95% lower CI bounds. Those values are also small.

Though with the estimated low benefit of the end-number license plate driving restriction in air pollution reduction, cost side of the policy is ambiguous and harder to evaluate. However, many studies have suspected potential high costs of this long-run version of road rationing. A survey study in Tianjin has suggested high public objection to the policy (Jia etal., 2017), and constant arguments have arisen regarding property rights, arguing against the legibility to ban private vehicles on road in the long term (Qian, 2011; Zhang, 2015). In addition, there are implications of the driving restriction delaying people's time to work, effectively reducing labor time and economic outputs (Viard and Fu, 2015). On the benefit side besides air pollution reduction effects, the policy may also have its gain in congestion release (Sun, Zheng and Wang, 2014; Fan etal., 2017).[19] Other studies showing transportation being a main driving force of carbon emission (Wang and Liu, 2015) have also advised on applying the driving restriction to reduce CO_2 emission, which is closely related to the topic of climate changes. With these potential direct or indirect costs and benefits, one may not easily reach the conclusion that this driving restriction is not obtaining any social

[18]Source: https://aqli.epic.uchicago.edu/about/methodology/

[19]A time-use dataset (https://www.cpc.unc.edu/projects/china/data/datasets/longitudinal/datasets) managed by Chinese CDC and the University of North Carolina is being requested for future research of this project. The plan is to analyze the policy effect on people's transportation time as well as labor time to estimate the potential cost and benefit in traffic and labor addressed in these literatures.

welfare gains. However, the main results showing relatively small benefits from air pollution reduction still address that one main goal implementing this policy have not been efficiently achieved.

3.6.2 Simultaneous Air Pollution Controls

One confounding factor in my driving restriction policy effect analysis remains that the policy may be compounded by other important environmental policies executed at the same time, thus the estimate on driving restriction will be representing actually a compounding effect of those simultaneous policies. This scenario is possible, since environmental policies have been fast to propose and execute in China over the past decades due to deteriorating pollutions, and it has been noted that governments are likely to take more than one measure in order to lower their city level air pollutions. But on the other hand, the end-number license plate driving restriction is likely a special policy managed by the local transportation department and applies to the most population unlike the other environmental policies usually imposed on polluting industries. It is also a policy not involving longer time-line costly monitoring, fining and negotiating between government and corporates, so it is more likely to be implemented fast and take effect swiftly.

To briefly explore whether there are more restricting simultaneous environmental regulations on air pollution in treatment cities around the same time as the driving restriction, I compile same set-up DID regressions to analyze city-annual level industrial pollution measures obtained from China City Statistical Yearbook. This exercise is conducted since many

environmental policies have been targeting heavy polluting industries, so if driving restriction comes in a package with these industrial level policies imposed by the local government, industries in treatment cities are likely also facing stricter regulations regarding their air pollutant emissions around the same time. I include the set of covariates of aggregated annual daily average temperature and precipitation, population, population density, primary, secondary and tertiary industrial share of GDP, GDP per capita converted to 2015 value, unemployment rate, private and self employment rate, ratio of labor force to city population, ratio of number of cell phone registers to city population, ratio of number of Internet registers to city population, and road area per capita. I translate the policy dummy to yearly level by defining year t under policy if at least half of t (six months) is influenced under the policy. For example, for Beijing executing the policy October 2008, year 2009 is the first year with $Policy_{it} = 1$.

For a couple of industrial pollution measures, the impacts of the road rationing policy timings are presented under **Table 3.G.1** for five-nearest-neighbor-matched control group. From the result table, we see there little statistical significance across the industrial pollution measures, representing little concerns of simultaneous environmental policies targeting at industries. However, though the water and solid waste treatment results can be less of a concern since they are less likely to interact with air pollution, industrial dust and SO_2 emissions will still be relevant. The estimates imply that policy cities may be having only 2.6% less dust emission and 14.3% more SO_2 emissions, the latter 10% significant statistically. Instead of implying a stricter industrial level air pollutants emission controls, these two estimates in turn present the possibilities of slacker industrial regulations in industries. Since

they imply more contribution from industrial emissions for policy cities post-treatment, the main conclusion in previous section may be altered since the long run end-number license plate driving restriction may contribute more to air quality improvement, but the increased industrial emissions undo the changes. But since there is not enough evidence and documentations about the treatment cities have more polluting industries, and this set of results are with low statistical significance and sensitive to identification strategies[20], I will put less weight on these arguments.

3.6.3 Effects on Transportation and Traffic

The government of China has been persuading the public on the negative effects of private car emissions deteriorating air pollution over the years, and they have been encouraging and promoting alternative public transportations[21]. Actually, the driving restriction policy only imposed on private vehicles is usually also regarded as one of these measures encouraging public transportations. As a result, we may expect public transportation and traffic variables to change accordingly in the long run due to the policy change. To test that, with the same approximation and covariates as previous subsection, city level annual transportation and traffic variables are regressed with DID. Results are presented under **Table 3.G.2**.

Overall across the panels estimates are again with low statistical significance, but that may be contributed by power issues due to low number of annual observations. It appears that among the number of unregulated vehicles (buses and taxis), volume of passengers by

[20]Actually if regressions are run no-logs instead of logs, the sign of these estimates will invert.
[21]Source: http://www.gov.cn/fwxx/kp/2007-09/17/content_751352.htm.

buses, number of buses, number of buses per population, and total passenger traveled per bus all have small and statistically insignificant estimates, while number of taxis has statistically insignificant 5.6% increase in policy cities relative to control cities. Number of passengers on highway has increased but insignificantly statistically. Therefore, there has been little evidence that public transportation like buses have been encouraged in the treatment cities both after the restriction applies on private vehicles. Instead, unregulated small vehicles that likely are equally polluting as private vehicles such as taxis may have increased. That might be one of the reasons why the driving restriction is not working as effectively as expected in long-run air pollution reduction. But still there is lack detailed data for individual travel time as well as actual traffic conditions (e.g., average speed of traveling within the city, number of car accidents), these results on transportation and traffic will remain relatively inconclusive.

3.6.4 Automobile Ownership

One of the argument why the driving restriction is not improving long run air quality is that people will buy more cars with different license plate end numbers in so the policy effect on air quality diminishes over time (Ma and He, 2016). In China, all vehicles must be registered with the local transportation department before getting a plate by lottery. In case this argument holds, we will expect to see more private vehicles registered on road in treatment versus control cities post-policy. To explore briefly this argument, I obtain the province-by-year private vehicles ownership dataset provided by China Auto Market Almanac and regress log of automobile ownership variables with the usual DID design. Since I

lack city level data, I have to extend policy treatment-control status to province level, which effectively reduce the number of controls making this set of analysis only proxies to previous city level results. But since all the policy cities are capitals of the respective provinces where population density is highest and GDP per capita the greatest, this approximation would still be somewhat reasonable and instructive.

Results are presented in **Table 3.G.3** broken down into subcategories of automobile types classified by the Almanac (for passenger, cargo, other use, among which passenger use takes the majority share), with province level population, population density, population weighted average primary, secondary and tertiary share, GDP per capita converted to 2015 value, unemployment rate, ratio of labor force to city population, and road area per capita as covariates[22]. From this result table, we see a total around 3.5% decrease in automobiles owned, but such estimate remains statistically insignificant, unable to verify the policy actually affecting number of private vehicles owned. We also see the coefficients are mainly negative, which may be attributed to decrease in private car ownership due to the policy, contrary to the prediction that citizens will purchase more cars to avoid the driving restriction. Note however, the test is again underpowered and biased due to approximation to province level.

The same DID analysis is conducted for the yearbook provided annual dataset of new registered civil use automobiles, which can be regarded as a proxy for yearly auto sales.

[22]Since number of observations decreased when panel is by province, I reduce number of covariates that are less likely to correlate with auto markets.

Note civil use autos would include all the private vehicles but also those purchased by non-governmental groups such as schools and corporates, though they not classified as private cars will report to local governments and not subject to the end-number license plate driving restriction. Results are presented under **Table 3.G.4**, from which there is again no statistically significant estimates showing the policy affecting new car purchases in treatment provinces. However point estimates are larger, with 17.2% decrease of newly registered passenger cars while the other types have positive changes in new registrations. Again, a decrease in passenger cars (which takes the majority of all civil use autos) is contrary to the argument that policy citizens purchase more cars to avoid the policy.

3.6.5 Traffic and Auto Market Changes by Individual Cities

From previous section, I have found certain empirical evidence by synthetic control method showing different cities with heterogeneous efficiency of the policy in air pollution reduction. To entangle the mechanisms and reasonings behind the set of heterogeneous effects, I conduct some preliminary DID analysis for individual treatment cities showing significant and large estimates in the synthetic control DID results **Table 3.F.1** on selected dependent variables of traffic and auto market variables presented in previous subsections. The 5 treatment cities with statistically significant estimates of log AQI regression under synthetic controls are Guiyang and Hangzhou with negative significant estimates, and Chengdu, Harbin and Tianjin with positive significant estimates. All of those estimates have magnitudes rank around 10%, with Tianjin higher above 20%. For this set of analysis, control groups include all cities/provinces available without the policy treatment, and treatment

city is only the one specified.[23].

Results are presented in **Table 3.G.5**. We see the two treatment cities with negative policy effects on AQI in previous section, Guiyang and Hangzhou, have reduced number of buses and travels by buses, while the number of taxis increases like in the pooled analysis **Section 6.3**. Both provinces of the capitals have little to a significant decrease in private cars ownership. On the other hands, among the three cities with positive significant policy effects on AQI, Chengdu has number of buses, number of travels by buses and number of taxis all increase significantly while private vehicles number decreases significantly, contrary to the belief that the policy works counter intuitively since people buy more private cars but agree with the argument that passengers switch to public transportation. The other two cities, Harbin and Tianjin both have some decrease in public transportation, buses for Harbin and taxis for Tianjin. They also show decrease in number of total private vehicles, Tianjin with a large magnitude and statistically significant estimate, and Harbin smaller and insignificant from zero.

These results of breakdown analysis by cities have again shown that purchasing more private cars to avoid the policy intervention is unlikely to intervene with the driving restriction policy as almost all selected policy cities have negative estimates on private owned vehicles post-policy. It also cannot confirm that more public transportations through the policy effects, either buses or taxis, are good or bad for the air qualities of the cities. For example, both Chengdu and Tianjin have positive policy impacts on AQI but the former

[23]Alternative designs such as synthetic control is not pursued at this stage because insufficient observations.

simultaneously experience increase in both buses and taxis, while the latter experiences a decrease. For another example, Hangzhou and Harbin both have decreased buses and increased taxis by the policy, but the former city have negative policy effect on AQI and the latter has positive. Those results may again turn to the reasonings that public transportations not subject to the driving restriction potentially can contribute greater to city air pollution. Since cities adopting to long-run end-number license plate control policies are capital cities with great incentive and government subsidies to provide clean public transportations such as electric buses[24], this argument can be quite counterintuitive. Again note the reasonings are not very conclusive due to power issues and the designs relatively biased using full set of control cities/provinces.

3.6.6 Effects on Labor Time

As another main purpose of the end-number license plate driving restriction is to reduce traffic congestions, an additional measure of the effectiveness of the policy will be whether it sufficiently reduce traffic congestion, increase convenience of travels within the cities of policy treatments, and positively affect people's labor activities. For analysis of that purpose, the China Health and Nutrition Survey (CHNS) dataset is obtained and identified at city level (ADM2). This dataset include around $30,000$ individuals from $7,200$ households stratified randomized selected from 15 provinces of China from 1989-2015 with about by-yearly intervals. Surveys are usually conducted second half of the year August to December. For the purpose of this project, I restrict to surveys between 2005 and 2015, which includes 2006,

[24]Source: https://kuaibao.qq.com/s/20190817A060OU00?refer=spider.

2009, 2011 and 2015.

Individual interview questions include time traveled to work, however is not yet compiled in their merged dataset. Instead, the alternative time-use variable this project is going to look at for policy impact is labor time previous week of the interview. Policy intervention is defined as if the whole week where labor time indicates falls after the policy change (which is to say, the Monday of the "last week" of the survey date will be already under the driving restriction policy). Due to sample restriction, there are only three treatment cities included in the CHNS dataset, Beijing, Harbin and Guiyang. Among them since Beijing enters the survey after its policy in 2008, only Harbin and Guiyang provide time variation to identify the policy effect on labor time through a DID design. As a result, the conclusion of this section will not be quite conclusive to extrapolate to the overall treatment effect of the policy because two treatment cities are unlikely to represent the other five.

Due to limited sample size, all cities without policy change during the periods have been included as the full control set in the DID following. Runs are also performed limited to urban surveyors, rural surveyors and only to narrow window 2007-2012 (thus only include 2009 and 2011 two surveys). For covariates, temperature and precipitation are summed up to weekly level to match the frequency of the data. Following the main DID design, city-by-year fixed effects and week fixed effects are added. Other covariates likely determinant of individual labor time have been included, indicator of gender, nationality, married or not, education levels, categories of primary occupation, whether employed or not, whether living in urban area, age, age squared, and household size. Since those covariates come from dif-

ferent survey parts of CHNS, including them would notably decrease the sample size in the regression. Again standard errors are clustered at city-by-year levels.

The regression results are presented under **Table 3.G.6**. From which we can identify a consistent significant negative impact of the driving restriction on labor time. For larger sample without covariates, the impact is 7% to 11% reduction of labor time no matter residents of urban or rural areas of the city. For smaller sample with covariates added, the impact is as high as 30%-47% labor time reduction. Within the narrow window, the effect is smaller, insignificant 5% without covariates and 26% with covariates, which may indicates a growing effect on labor time reduction by the policy change.

Though previous argument on if policy reduce congestion time and decrease travel time may be increasing labor time is counter-proved by these regressions, there exist reasonable explanations. Though congestion times may be reduced, travel time may still increase for most workers who cannot afford to buy a second car avoiding the policy intervention preventing them to drive for work. That will make people using public transportation for work, possibly reducing their labor time either due to selection not going because of inconvenience of traveling to work, or otherwise labor time is replaced by longer travel time. Another less possible explanation is though the survey asks explicitly labor time, many surveyors may still regard travel time to work as part of the time for work though it is not being paid. In that case the results actually may indicate a reduction of travel time due to possible congestion release in the cities of policy treatments.

If we evaluate the results from this extension section seriously, we may be reaching the conclusion that the driving restriction policy not only did not sufficiently improve air quality in the long run, but also drive down labor time in treatment cities likely implying a counter effect on people's daily convenience of traveling around the city. That will then add to cost side of the policy and reduce its net benefit. However note that this exercise is inconclusive due to limited sample size only covering two of the nine treatment cities, and their may be multiple errors in the survey questions, so such increase in policy cost is likely still be uncertain.

3.7 Robustness Check

In this section, I will run some robustness checks for the main DID result **Table 3.3**. First of all, since air quality records coming from ground monitoring stations are usually under speculations of whether data is truthfully reported, I will conduct analysis on the main AQI dataset excluding observations suspicious of data manipulations. Then, I will run regressions with alternative air pollution measures. According to Zhang, Lin Lawell and Umanskaya (2017), different pollutants can be impacted differently by policies, therefore my conclusion from AQI index alone may not be extended to the scope of all pollutants. Moreover, the old standard AQI I take in the main analysis have ignored some major pollutants of concern like *PM2.5*. Therefore, I will run robustness checks on individual pollutants including *PM2.5* and *PM10*. Lastly, I will also repeat main analysis with alternative data source from satellite images, the AOD index. Though these robustness checks may cast some

doubts on the validity of my main results, note none of the alternative data sources have comparable frequency and coverage as the main AQI dataset, which is a reason why I will still regard the main AQI result with more weights.

3.7.1 Suspects of Data Quality Change

The reform in 2012 has not only change standards evaluating AQI in China, but also started the process of making air quality real time, public and automated[25]. The change to automation may cause a sudden jump in reported AQI levels, because manipulation have to drop after automation happens. Though I could not identify when automation happens in each city and whether manipulation exists before the process, the automation always happens on January 1st per year. Therefore, I use regression discontinuity in time (RDiT) to estimate the jump on reported AQI for each individual cities across the threshold of January 1 each year 2013-2015. From those RDiT results, I classify any cities with at least 10% statistically positive significant jumps across January 1st, for all three bandwidth windows $30, 180, 365$ polynomial order 1, as cities with data manipulations before automations. That will leave me with 64 cities, with one of the treatment city, Chengdu, also dropped from the sample. I then repeat DID on this set of cities excluding those with suspicious data manipulations, and results are presented under **Table 3.H.1**. Comparing with the main results **Table 3.3**, both negative point estimates and lower 95% CI bounds have increased in magnitude, though all still remains statistically insignificant from zero once adding covariates. Estimates shows at most 1.5 unit decrease of AQI by point estimates, and most negative maximum

[25]Reference with draft of Professor He's working paper, "Can Technology Solve the Principal-Agent Problem? Evidence from Pollution Monitoring in China", with Michael Greenstone, Ruixue Jia, and Tong Liu.

reductions of 8.1 by 95% confidence interval lower bounds. These numbers are not huge but still considerably larger than the main results. Since one of the treatment cities with positive estimate in synthetic control section has been dropped, the outcome is not surprising.

In addition, more monitoring stations are added across China since start of 2013, leading to large increase in the number of cities covered in my main AQI dataset from above 100 to above 300 during 2013-2014. Even in the sample cities I have selected with observations throughout the sample period, the number of monitoring stations where the daily city level AQI is averaged across has increased in time, giving possibilities to sudden shift in data quality. Since AQI measurement will be much different from before or after the reform, I perform the main DID with a narrow window of 2007-2012 restricting to prior to the reform. Within this time range cities with the policy treatment decrease to 7, excluding Lanzhou and Tianjin. **Table 3.H.2** presents this set of results, showing generally zero and even positive insignificant estimates of the policy, much similar to the main table. Unlike the main results, even specifications without covariates have been small in magnitude and statistically insignificant, verifying the general zero effect of the policy at least until 2012.

3.7.2 Results with PM2.5

I obtain the US embassy hourly measurement of *PM2.5* concentration for 5 cities, Beijing, Chengdu, Guangzhou, Shanghai and Shenyang for years 2010-2015, within which two (Beijing and Chengdu) have the policy treatment. I repeat DID for this dataset adding policy indicator interacting with a peak hour dummy (7am-8pm), which are the hours the policy

is in effect. Results are presented under **Table 3.H.3**. Negative insignificant estimates on the interactive dummy variable are identified throughout specifications. Magnitudes of these point estimates are not huge comparing to the city average *PM2.5* of Beijing and Chengdu, $91.3\mu g/m^3$, which counts to about 8% reduction and up to 22% at lower CI. This is still notably larger than the AQI reduction in my main specification.

Since my AQI measure does not consider *PM2.5*, this set of results are instructive on the potential higher benefit of the driving restriction in air quality improvement through lowering *PM2.5* concentrations. However, these results shall be treated with caution. Unlike AQI, the measure of *PM2.5* has not been with great familiarity among public until 2012. and there maintains various debates across China regarding whether car emissions contribute more or less to the *PM2.5* concentration in the city without very conclusive evidence[26]. Also the *PM2.5* dataset is quite limited in number of city and years covered, though it is very high frequency hourly. It comes from a monitoring source outside of China and sometimes more trusted for less data manipulations, but the US Embassy only have one monitoring point each city so the measurement can also be less representative.

To value the effect on the same scale as AQI, I use conversion factor to *PM10* from EPIC, which is 0.65 *PM2.5*-to-*PM10*. That then gives approximate equivalent $11\mu g/m^3$ peak hour *PM10* reduction by this long-run policy, equivalent to a much larger value gain of 50.93 Yuan per person per year, maximum using 95% CI lower bound being 145.51 Yuan per person per year, halving the effect since only applying to peak hours of a day. This result, though

[26]Source: http://news.ifeng.com/a/20170103/50512431_0.shtml.

doubled comparing with the main AQI results, is still affordable comparing with annual per capita household income in most urban areas of China.

3.7.3 Results with PM10

With the annual daily average *PM10* pollution data 2005-2014 from Professor Greenstone's multiple studies, I rerun my DID with annual average daily *PM10* as the dependent variable. Since the data is annual by city I again use the approximations in **Section 6** to extend my policy treatment dummy from daily to yearly, and include the same city-by-year level covariates controls from China City Statistical Yearbook. Note like all regressions in **Section 6**, these regressions are underpowered comparing with other datasets I have been using in this section. Estimates from these regressions are also more subject to omitted variable biases not included in the yearbook source, like unobserved auto market changes and unobserved consumer preference shocks.

Regression results with covariates are shown in **Table 3.H.4**. To summarize, there is a sizable effect of the policy for 7.0–10.4$\mu g/m^3$ *PM10* reduction except for the only policy cities regression **Column (1)**. Statistical significance remains low, only 10% with same-province and full control cities. The magnitude is not small but also not huge again comparing with treatment cities pre-policy mean of 98$\mu g/m^3$, which counts to 10%-23% reduction maximum at lower 95% CI bounds. Translating to WTP directly, that corresponds to about 13.27 – 96.64 Yuan per year per person gains (maximum 89.52 – 210.62 at 95% CI). This is considerably greater than the AQI and *PM2.5* estimates, though still affordable and not

super huge numbers.

3.7.4 AOD

One other popular source identifying air pollution is by satellite image, which is free from ground monitoring station mistakes and manipulations. The AOD measure (Aerosol Optical Distance) measured by NASA provides this alternative. However it simultaneously possesses other disadvantages comparing with ground station records, one being besides air pollutants, AOD is also related to confounding factors of windblown dust, sea salts, volcanic ash, smoke from fires, etc.[27].

Since satellite cannot pass each location on earth every single day, the AOD data is usually taken monthly. The AOD variable is recorded on the scale of $[0, 1]$. I downloaded 0.1×0.1 grid resolution monthly AOD 2005-2015 from NASA and compile data to city level, through matching each city centroids to 4 closest 0.1×0.1 grids, then computing the city average AOD as the weighted average of the 4 grids AOD by inverse distance square[28]. I convert the policy indicator used in DID design again to monthly level, defining at least half a month falling under the policy as a policy treated month. I repeat my DID analysis with monthly aggregate daily average temperature and precipitation as covariates, and again adding city-by-year FE.

Results are shown under **Table 3.H.5**. Point estimates are again small and never statistically significant. However the mean AOD for treatment cities pre-policy is low around

[27]Source: https://neo.sci.gsfc.nasa.gov/view.php?datasetId=MODAL2_M_AER_OD.
[28]Methods of matching referencing He, Fan and Zhou (2016).

0.46, thus the point estimates can reach as large as 14% reduction, and maximum 38% by lower 95% CI bound. We may call these results noisy and uncertain, not contradicting with the main results, but there is still possibilities there is a sizable air pollution reduction effect from the policy based on AOD measurements.

3.8 Conclusion

To summarize, this project has suggested little to no long run average effects improving city level air quality of the continuous version Chinese end-number license plate driving restriction, across all 9 cities implementing the policy in long-term 2008-2013. Furthermore, policy may be effective and ineffective for different cities, in immediate short-run and long-run. Besides air quality changes, there have also not been found any effects of the policy on traffic, transportation and private automobile ownership.

There could be various explanations why the policy is actually not on average effective in air quality improvement suggested by previous literatures. One of the popular explanation is quick behavior response, such as people buying more cars with different license plate end numbers so total number of cars on roads do not change over the long run (Ma and He, 2016). Others have suggested improvement of traffic condition accompanied by the policy instead would increase air pollution, as traffic and pollution have non-monotonic correlation (Sun, Zheng and Wang, 2014). There are also arguments that the policy is not enforced strict enough (Lu, 2016), such that if with higher fines and stricter rules on purchasing a second car, the driving population will be able to conform by driving less and then leading

to fewer car emissions and less air pollution.

From my analysis, I also tend to believe the policy is not effective enough even in the short run as found in the narrow window DID and the RDiT design. For the reasons and mechanisms of these results, through my extension regression analysis, I do not think it is because citizens purchasing a second private car and changes in public transportation is evident of the policy response. This is probably consistent with intuitions, since affording a second car is not really universal in most Chinese households, let alone there have been strict rules regarding how people are randomly assigned a license plate number of their new cars in these cities. That potentially put on high enough cost of buying and supporting a new car with the risk of having a plate end number same group of the current one, comparing with inconvenience of not using private cars one-fifth of the week. However, my results also imply that when private cars are forbidden from roads, passengers may switch more to transport by taxis instead of buses, and that indirect effect on air pollution remains ambiguous. In addition, I am also less convinced on the explanation that the restriction is not applied strictly. The policy is probably binding for the public, since it has long been under much debates after over a decades, and most drivers knowing about the increasing level of surveillance on road asks and cares about being caught as a violator of the policy. Therefore, I inclined to think that the policy is not that effective in controlling for air pollution either because car emissions contribute little to air pollutions of the city, or the policy decreases the number of cars on road but maintains level or car emissions by putting more public transportations or enabling current cars driving faster.

With the ambiguous and on average little effect of air pollution reduction by this policy, together with potentially higher cost of implementation, it is interesting to think of why local governments still try to implement the policy for long run. One reason argued is that this policy is chosen and favored because it can be easily and swiftly implemented across the city, and it is associates with relatively lower cost for the local governments comparing with other environmental policies usually involving monitoring and regulating high polluting industries. It is suggestive that policy makers have political concerns when public get annoyed by continuous air pollutions in the city, so that they may have the urgent need to take and stay with this progressive form of policy such that it appears an effort has been made to control air pollution (Zheng etal., 2014). Furthermore, as the driving restriction policy gets more and more popular across the country because they are reported to effectively release short run air pollution problem, these governments may find it straightforward extrapolating the short-run strict form of the driving restriction to the long-run. However, all these explanations will still require more theoretical grounds and empirical evidence, since they imply that myopic governments are making such decisions and do not correct accordingly in time.

3.9 References

Carrillo, Paul E., Arun S. Malik, and Yiseon Yoo. "Driving restrictions that work? Quito's Pico y Placa Program." *Canadian Journal of Economics* 49, no. 4 (2016): 1536-1568.

Chen, Yuyu, Ginger Zhe Jin, Naresh Kumar, and Guang Shi. "The promise of Beijing: Evaluating the impact of the 2008 Olympic Games on air quality." *Journal of Environmental*

Economics and Management 66, no. 3 (2013): 424-443.

Cicala, Steve. "When does regulation distort costs? lessons from fuel procurement in us electricity generation." *American Economic Review* 105, no. 1 (2015): 411-44.

Currie, Janet, John Voorheis and Reed Walker. "What Caused Racial Disparities in Particulate Exposure to Fall? New Evidence from the Clean Air Act and Satellite-Based Measures of Air Quality." *Working Paper*, 2019.

Davis, Lucas W. "The effect of driving restrictions on air quality in Mexico City." *Journal of Political Economy* 116, no. 1 (2008): 38-81.

Fan Shoubin, Tian Lingxi, Guo Jinjin, and Sun Gaihong. "Effect of Odd-Even Traffic Restriction on Exhaust Emission of Suburban Highway." *Journal of Environmental Engineering Technology* 7, no. 5 (2017): 539-545.

Ge Keyou, *Health and Nutrition Status of Residents in Eight Provinces of China*, Vol. 1. Beijing Science and Technology Press, 1998.

Jia, Ning, Yidan Zhang, Zhengbing He, and Geng Li. "Commuters' acceptance of and behavior reactions to license plate restriction policy: A case study of Tianjin, China." *Transportation Research Part D: Transport and Environment* 52 (2017): 428-440.

Hausman, Catherine, and David S. Rapson. "Regression discontinuity in time: Considerations for empirical applications." *Annual Review of Resource Economics* 10 (2018): 533-552.

He, Guojun, Maoyong Fan, and Maigeng Zhou. "The effect of air pollution on mortality in China: Evidence from the 2008 Beijing Olympic Games." *Journal of Environmental Economics and Management* 79 (2016): 18-39.

Huang, Hengjun, Deyin Fu, and Wei Qi. "Effect of driving restrictions on air quality in Lanzhou, China: Analysis integrated with internet data source." *Journal of cleaner production* 142 (2017): 1013-1020.

Ito, Koichiro, and Shuang Zhang. Willingness to pay for clean air: Evidence from air purifier markets in China. No. w22367. National Bureau of Economic Research, 2016.

Lu, Xueying. "Effectiveness of government enforcement in driving restrictions: a case in Beijing, China." *Environmental Economics and Policy Studies* 18, no. 1 (2016): 63-92.

Ma, Hua, and Guizhen He. "Effects of the post-olympics driving restrictions on air quality in Beijing." *Sustainability* 8, no. 9 (2016): 902.

Qian Qing. Interpretation of Administrative Law on Driving Restrictions: A Case of Odd-Even Rationing. *Administrative Law Research* vol. 4 (2011).

Sun, Cong, Siqi Zheng, and Rui Wang. "Restricting driving for better traffic and clearer skies: Did it work in Beijing?." *Transport Policy* 32 (2014): 34-41.

Viard, V. Brian, and Shihe Fu. "The effect of Beijing's driving restrictions on pollution and economic activity." *Journal of Public Economics* 125 (2015): 98-115.

Wang, Zhaohua, and Wei Liu. "Determinants of CO2 emissions from household daily travel in Beijing, China: Individual travel characteristic perspectives." *Applied Energy* 158 (2015): 292-299.

Xu Gonghu, and Hou Jiayin. "Study on the Impact of Motor Vehicle Restriction Policy on Air Quality: Taking Chengdu as an Example." *Public Economics and Policy Research* 1 (2015): 9.

Ye, C., and N. Zhuo. "How could local government s policies improve air quality?-Empirical analysis to check local government s policies to deal with air pollution in Hangzhou, China." (2018).

Zhang Xiang. "Driving Restriction, Property Rights and Proportional Principles." PhD diss., 2015.

Zhang, Wei, C-Y. Cynthia Lin Lawell, and Victoria I. Umanskaya. "The effects of license plate-based driving restrictions on air quality: Theory and empirical evidence." *Journal of Environmental Economics and Management* 82 (2017): 181-220.

Zheng, Siqi, Matthew E. Kahn, Weizeng Sun, and Danglun Luo. "Incentives for China's urban mayors to mitigate pollution externalities: The role of the central government and public environmentalism." *Regional Science and Urban Economics* 47 (2014): 61-71.

3.A Air Quality Measurement Standards

Figure 3.A.1: Air Quality Index Conversion Table, New Standard GB 3095-1996

（一）空气污染指数的定义及分级限值

API（Air Pollution Index 的英文缩写）是空气污染指数，我国城市空气质量日报 API 分级标准如表 1：

表 1 空气污染指数对应的污染物浓度限值

污染指数	污染物浓度（毫克/立方米）				
API	SO_2（日均值）	NO_2（日均值）	PM_{10}（日均值）	CO（小时均值）	O_3（小时均值）
50	0.050	0.080	0.050	5	0.120
100	0.150	0.120	0.150	10	0.200
200	0.800	0.280	0.350	60	0.400
300	1.600	0.565	0.420	90	0.800
400	2.100	0.750	0.500	120	1.000
500	2.620	0.940	0.600	150	1.200

表 2 空气污染指数范围及相应的空气质量类别

Figure 3.A.2: Air Quality Index Conversion Table, New Standard GB 3095-2012

HJ 633—2012

表 1　空气质量分指数及对应的污染物项目浓度限值

空气质量分指数（IAQI）	污染物项目浓度限值									
	二氧化硫（SO_2）24 小时平均/（μg/m³）	二氧化硫（SO_2）1 小时平均/（μg/m³）(1)	二氧化氮（NO_2）24 小时平均/（μg/m³）	二氧化氮（NO_2）1 小时平均/（μg/m³）(1)	颗粒物（粒径小于等于10μm）24 小时平均/（μg/m³）	一氧化碳（CO）24 小时平均/（mg/m³）	一氧化碳（CO）1 小时平均/（mg/m³）(1)	臭氧（O_3）1 小时平均/（μg/m³）	臭氧（O_3）8 小时滑动平均/（μg/m³）	颗粒物（粒径小于等于2.5μm）24 小时平均/（μg/m³）
0	0	0	0	0	0	0	0	0	0	0
50	50	150	40	100	50	2	5	160	100	35
100	150	500	80	200	150	4	10	200	160	75
150	475	650	180	700	250	14	35	300	215	115
200	800	800	280	1 200	350	24	60	400	265	150
300	1 600	(2)	565	2 340	420	36	90	800	800	250
400	2 100	(2)	750	3 090	500	48	120	1 000	(3)	350
500	2 620	(2)	940	3 840	600	60	150	1 200	(3)	500
说明：	(1) 二氧化硫（SO_2）、二氧化氮（NO_2）和一氧化碳（CO）的 1 小时平均浓度限值仅用于实时报，在日报中需使用相应污染物的 24 小时平均浓度限值。 (2) 二氧化硫（SO_2）1 小时平均浓度值高于 800 μg/m³ 的，不再进行其空气质量分指数计算，二氧化硫（SO_2）空气质量分指数按 24 小时平均浓度计算的分指数报告。 (3) 臭氧（O_3）8 小时平均浓度值高于 800 μg/m³ 的，不再进行其空气质量分指数计算，臭氧（O_3）空气质量分指数按 1 小时平均浓度计算的分指数报告。									

3.B Balance Checks

Table 3.B.1: Balance Table Check for Treatment and Control Cities, Average Over Period 2005-2007

Variable	Treatment Group Average	Control Group Average	Difference	P-Value
Same-Province Control Groups				
City Area (km^2)	23163.022	21056.342	-2106.679	0.809
Daily AQI	82.940	71.282	-11.659	0.035
Cellphone/Pop.	0.727	0.496	-0.231	0.071
GDP per Cap. (2015 Yuan)	42886.128	29717.716	-13168.412	0.130
Internet/Pop.	0.166	0.083	-0.083	0.020
Labor/Pop.	0.319	0.163	-0.156	0.008
Pop. Density (pc/m^2)	428.838	349.781	-79.057	0.375
Population (10000 pc)	753.891	494.924	-258.966	0.036
Daily Precip. (mm)	2.360	2.811	0.451	0.427
Prim. Ind. Shr. (%)	6.397	13.938	7.541	0.012
Prop. Private Employed (%)	34.108	36.161	2.054	0.475
Road Area Per Cap. (m^2)	8.800	9.591	0.791	0.560
Second. Ind. Shr. (%)	45.603	47.305	1.702	0.682
Daily Avg. Temp. ($^\circ C$)	11.933	13.859	1.927	0.396
Ter. Ind. Shr. (%)	47.999	38.757	-9.242	0.019
Unemployment Rate (%)	3.067	3.611	0.543	0.305
Five-Nearest-Neighbor-Matched Control Group				
City Area (km^2)	23163.022	13961.654	-9201.368	0.185
Daily AQI	82.940	70.638	-12.302	0.023
Cellphone/Pop.	0.727	0.495	-0.232	0.028
GDP per Cap. (2015 Yuan)	42886.128	35140.719	-7745.409	0.257
Internet/Pop.	0.166	0.081	-0.084	0.003
Labor/Pop.	0.319	0.198	-0.122	0.033
Pop. Density (pc/m^2)	428.838	479.300	50.462	0.543
Population (10000 pc)	753.891	487.858	-266.033	0.019
Daily Precip. (mm)	2.360	2.439	0.080	0.864
Prim. Ind. Shr. (%)	6.397	10.507	4.110	0.027
Prop. Private Employed (%)	34.108	36.809	2.701	0.267
Road Area Per Cap. (m^2)	8.800	10.907	2.107	0.011
Second. Ind. Shr. (%)	45.603	51.485	5.882	0.071
Daily Avg. Temp. ($^\circ C$)	11.933	14.506	2.573	0.130
Ter. Ind. Shr. (%)	47.999	38.008	-9.991	0.003
Unemployment Rate (%)	3.067	3.345	0.278	0.436

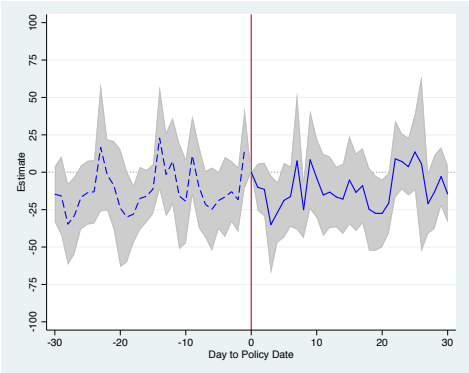
Table 3.B.1, continued

Variable	Treatment Group Average	Control Group Average	Difference	P-Value
		Full Control Group		
City Area (km^2)	23163.022	16690.159	-6472.862	0.326
Daily AQI	82.940	71.337	-11.604	0.024
Cellphone/Pop.	0.727	0.597	-0.130	0.562
GDP per Cap. (2015 Yuan)	42886.128	35947.585	-6938.543	0.407
Internet/Pop.	0.166	0.107	-0.059	0.282
Labor/Pop.	0.319	0.260	-0.059	0.538
Pop. Density (pc/m^2)	428.838	436.376	7.539	0.946
Population (10000 pc)	753.891	499.844	-254.047	0.070
Daily Precip. (mm)	2.360	2.548	0.189	0.647
Prim. Ind. Shr. (%)	6.397	10.218	3.821	0.130
Prop. Private Employed (%)	34.108	39.141	5.033	0.150
Road Area Per Cap. (m^2)	8.800	10.904	2.104	0.352
Second. Ind. Shr. (%)	45.603	49.522	3.919	0.283
Daily Avg. Temp. ($°C$)	11.933	14.507	2.574	0.173
Ter. Ind. Shr. (%)	47.999	40.259	-7.739	0.028
Unemployment Rate (%)	3.067	3.546	0.479	0.362

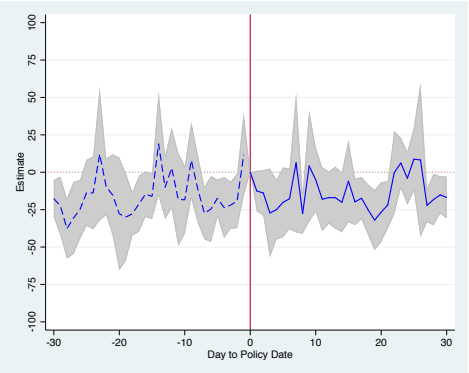
Note: Difference equals late group average minus early group average; P-value indicates T-test p-values between the group means; Annual city-level data source China City Statistical Yearbook; Daily AQI from Ministry of Ecology and Environment; ERAI climate data from Climate Impact Lab provides daily temperature and precipitation, aggregated from grid to prefecture level cities population weighted; GDP per capita in Yuan adjusted to 2015 by GDP deflator source World Bank.

3.C Event Study Supplementary Figures

Figure 3.C.1: Event Study with Window of 60 Days, Only Policy Cities

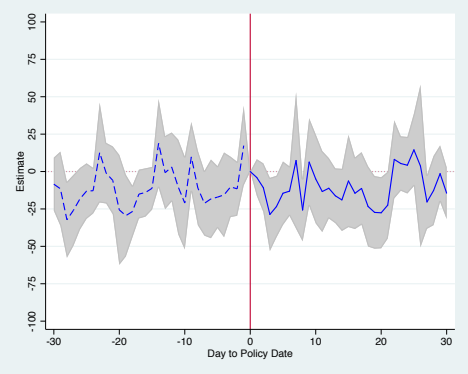


(a) Without Controls

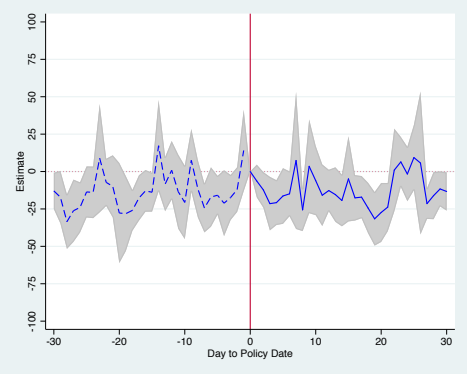


(b) With Controls

Figure 3.C.2: Event Study with Window of 60 Days, Controls within Same Provinces

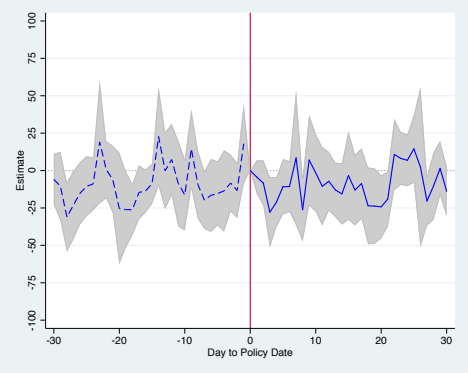


(a) Without Controls

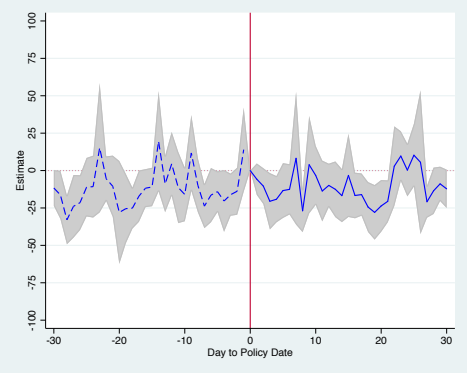


(b) With Controls

Figure 3.C.3: Event Study with Window of 60 Days, Matched Controls



(a) Without Controls



(b) With Controls

3.D Main Regression Supplementary Tables

Table 3.D.1: Difference-In-Difference (DID) Regression Results With Varying Matching Sample Size

	(1)	(2)	(3)	(4)	(5)	(6)
Policy	-7.773*	-0.084	-7.735***	-0.035	-6.423***	-0.177
	(4.055)	(2.057)	(1.880)	(1.972)	(1.687)	(2.103)
	[-15.784,0.238]	[-4.148,3.980]	[-11.431,-4.038]	[-3.912,3.841]	[-9.737,-3.109]	[-4.309,3.955]
Matched Size	1	1	5	5	10	10
Controls	No	Yes	No	Yes	No	Yes
SE	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered
N	54821	54821	136063	136063	178978	178978
Adj. R^2	0.314	0.459	0.326	0.449	0.329	0.441

Note: Standard errors in parentheses, clustering on city-by-year level; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is daily average AQI; Controls include climate variables (daily average temperature and daily precipitation), and indicators for major pollutants.

Table 3.D.2: Difference-In-Difference (DID) Regression Results With Varying Matching Sample Size, Alternative Matching Excluding Distance Measures

	(1)	(2)	(3)	(4)	(5)	(6)
Policy	-7.590***	-0.023	-6.267***	0.353	-5.106***	0.347
	(2.128)	(2.648)	(1.713)	(2.107)	(1.611)	(2.214)
	[-11.787,-3.392]	[-5.247,5.201]	[-9.634,-2.899]	[-3.789,4.495]	[-8.270,-1.942]	[-4.003,4.696]
Matched Size	1	1	5	5	10	10
Controls	No	Yes	No	Yes	No	Yes
SE	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered
N	66746	66746	151639	151639	190046	190046
Adj. R^2	0.353	0.463	0.333	0.443	0.333	0.438

Note: Standard errors in parentheses, clustering on city-by-year level; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is daily average AQI; Controls include climate variables (daily average temperature and daily precipitation), and indicators for major pollutants.

3.E Regression Discontinuity in Time Design by Cities

Table 3.E.1: Regression Discontinuity in Time (RDiT) Results, Individual Cities, Local Linear with Varying Bandwidths

Bandwidth	Beijing	Changchun	Chengdu	Guiyang	Hangzhou	Harbin	Lanzhou	Nanchang	Tianjin
30	0.176	-0.135	0.078	-0.455	0.286*	-0.264	-0.350*	-0.327***	-0.151
	(0.276)	(0.282)	(0.208)	(0.359)	(0.173)	(0.238)	(0.211)	(0.105)	(0.256)
180	0.424***	-0.125	-0.199**	0.009	0.264***	-0.031	-0.330***	-0.224***	0.081
	(0.131)	(0.088)	(0.092)	(0.117)	(0.080)	(0.089)	(0.096)	(0.062)	(0.103)
365	0.204**	-0.298***	-0.287***	0.057	0.209***	-0.210***	-0.292***	-0.067	0.172**
	(0.090)	(0.059)	(0.068)	(0.075)	(0.060)	(0.063)	(0.070)	(0.053)	(0.075)
Historical AQI	100.024	72.746	81.774	66.344	76.678	76.317	102.964	67.522	78.910

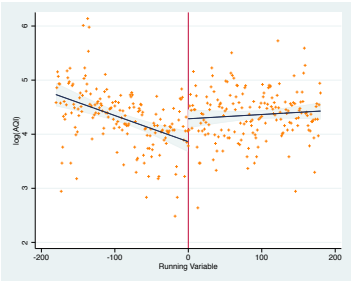
Note: Newey-West AR(1) standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is log(daily average AQI); Asymmetric local linear running variable (days to policy change) with a uniform kernel; Historical AQI is averaged across all days prior to the policy since 2005.

Table 3.E.2: Regression Discontinuity in Time (RDiT) Results, Individual Cities, Bandwidth 180 Days with Varying Local Polynomial Forms

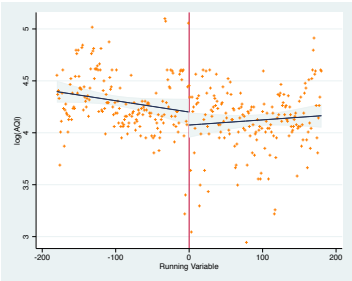
Poly Order	Beijing	Changchun	Chengdu	Guiyang	Hangzhou	Harbin	Lanzhou	Nanchang	Tianjin
1	0.424***	-0.125	-0.199**	0.009	0.264***	-0.031	-0.330***	-0.224***	0.081
	(0.131)	(0.088)	(0.092)	(0.117)	(0.080)	(0.089)	(0.096)	(0.062)	(0.103)
2	0.394*	0.019	-0.036	-0.056	0.156	0.194	0.002	-0.351***	-0.127
	(0.204)	(0.154)	(0.132)	(0.201)	(0.113)	(0.154)	(0.134)	(0.083)	(0.157)
3	0.080	-0.347	-0.098	-0.419	0.027	-0.271	0.059	-0.188*	0.016
	(0.261)	(0.220)	(0.169)	(0.268)	(0.152)	(0.189)	(0.180)	(0.099)	(0.202)
4	0.033	-0.265	0.379*	-0.407	0.167	-0.562**	-0.474**	-0.163	0.115
	(0.291)	(0.281)	(0.206)	(0.354)	(0.192)	(0.225)	(0.234)	(0.106)	(0.232)
Historical AQI	100.024	72.746	81.774	66.344	76.678	76.317	102.964	67.522	78.910

Note: Newey-West AR(1) standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is log(daily average AQI); Asymmetric local polynomials of running variable (days to policy change) with a uniform kernel; Historical AQI is averaged across all days prior to the policy since 2005.

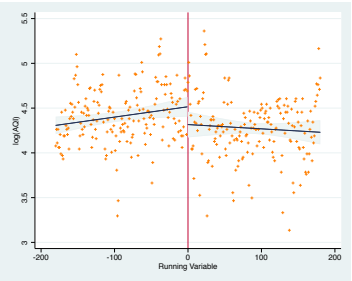
Figure 3.E.1: Regression Discontinuity in Time Design with $h = 180$, Individual Cities



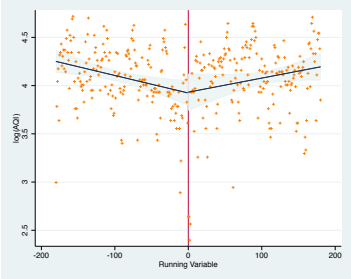
(a) Beijing



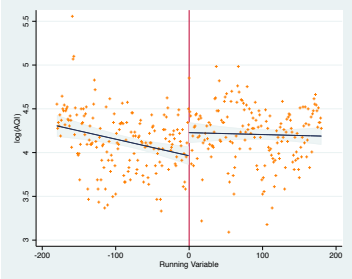
(b) Changchun



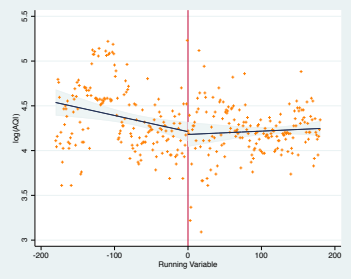
(c) Chengdu



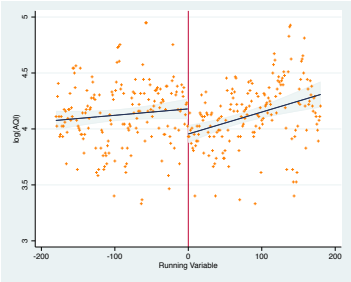
(d) Guiyang



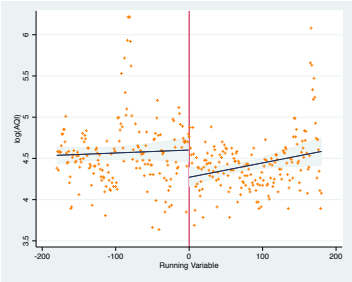
(e) Hangzhou



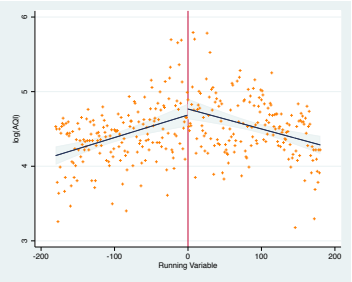
(f) Harbin



(g) Nanchang



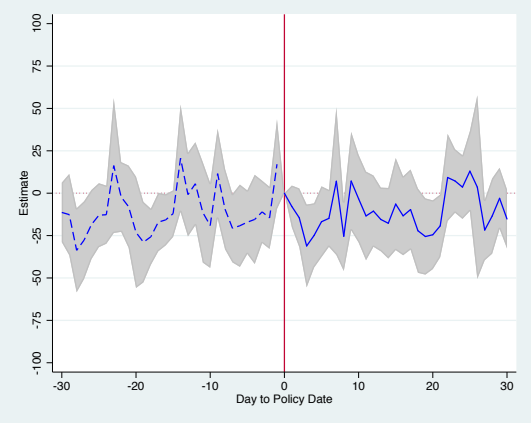
(h) Lanzhou



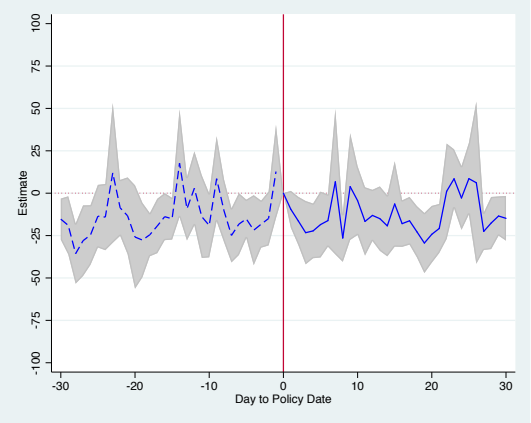
(i) Tianjin

3.F Synthetic Controls

Figure 3.F.1: Event Study with Synthetic Controls, Window of 60 Days, Full Controls

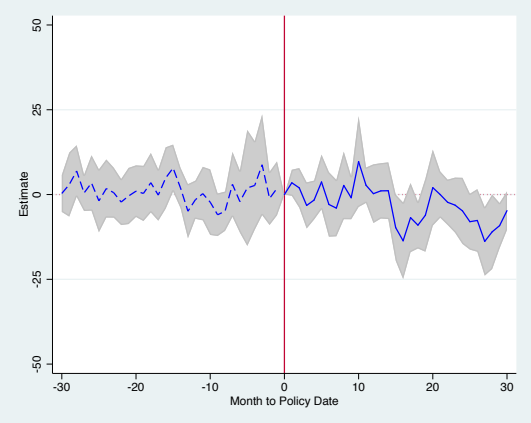


(a) Without Covariates

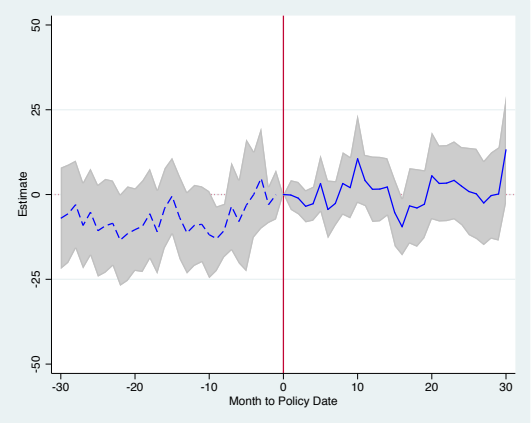


(b) With Covariates

Figure 3.F.2: Event Study with Synthetic Controls, Window of 60 Months, Full Controls



(a) Without Covariates



(b) With Covariates

Table 3.F.1: Difference-In-Difference (DID) with Synthetic Controls Regression Results

	Pooled	Beijing	Changchun	Chengdu	Guiyang	Hangzhou	Harbin	Lanzhou	Nanchang	Tianjin
Policy	0.008	-0.045	-0.029	0.137***	-0.094***	-0.110***	0.080***	0.069	0.021	0.236***
	(0.024)	(0.074)	(0.021)	(0.026)	(0.023)	(0.034)	(0.028)	(0.136)	(0.030)	(0.065)
	[-0.039,0.054]	[-0.194,0.105]	[-0.070,0.012]	[0.085,0.189]	[-0.140,-0.049]	[-0.177,-0.044]	[0.025,0.135]	[-0.203,0.341]	[-0.038,0.079]	[0.107,0.365]
Historical Avg. AQI	80.761	100.024	72.746	81.774	66.344	76.678	76.317	102.964	67.522	78.910
N	218330	19220	77093	43018	50432	47125	42273	19808	55139	47543
Adj. R^2	0.587	0.781	0.727	0.759	0.763	0.841	0.729	0.639	0.783	0.733

Note: Standard errors in parentheses, clustered by city-year; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is log(daily average AQI); Synthetic control weights constructed using pre-policy trends; Controls including daily average temperature and precipitation; FE includes city-by-year and date fixed effects; Historical AQI is averaged across all days prior to the policy since 2005.

3.G Extension Results

Table 3.G.1: Difference-In-Difference (DID) Regression Results, Extension on Industrial Pollution Emissions

	(1) Industrial Solid Waste Utilized (%)	(2) Waste Water Treated (%)	(3) Consumption Waste Treated (%)	(4) Industrial Dust Emission (10,000 tons)	(5) Industrial Waste Water Discharged (10,000 tons)	(6) Industrial SO_2 Emission (10,000 tons)
Policy	-0.263	-1.275	-10.952	-0.026	0.050	0.143*
	(2.667)	(3.139)	(7.221)	(0.103)	(0.080)	(0.073)
Control Cities	Matched	Matched	Matched	Matched	Matched	Matched
Controls	Yes	Yes	Yes	Yes	Yes	Yes
SE	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered
N	526	496	501	534	535	535
Adj. R^2	0.756	0.677	0.492	0.812	0.907	0.870

Note: Standard errors in parentheses, clustering on city level; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; City and year fixed effects are included; Dependent variables in logs for pollutant emissions and waste discharged; Controls include climate variables, demographic and macroeconomic variables.

Table 3.G.2: Difference-In-Difference (DID) Regression Results, Extension on Transportation and Traffic

	(1) Highway Passenger Traffic (10,000)	(2) Number of Buses	(3) Total Passengers Travel by Bus (10,000)	(4) Number of Taxis	(5) Number of Buses Per 10,000 Pop.	(6) Total Passengers/ Num. Buses (10,000)
Policy	0.317	-0.051	-0.005	0.056	-0.044	0.045
	(0.239)	(0.036)	(0.124)	(0.041)	(0.039)	(0.126)
Control Cities	Matched	Matched	Matched	Matched	Matched	Matched
Controls	Yes	Yes	Yes	Yes	Yes	Yes
SE	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered
N	529	536	534	536	536	534
Adj. R^2	0.834	0.985	0.918	0.982	0.901	0.464

Note: Standard errors in parentheses, clustering on city level; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Province and year fixed effects are included; Dependent variables all in natural logs; Controls include demographic and macroeconomic variables.

Table 3.G.3: Difference-In-Difference (DID) Regression Results, Extension on Private Owned Automobiles

	(1) Total	(2) Passenger	(3) Cargo	(4) Other
Policy	-0.035	-0.050	-0.059	-0.000
	(0.064)	(0.073)	(0.113)	(0.115)
Control Provinces	Full	Full	Full	Full
Controls	Yes	Yes	Yes	Yes
SE	Clustered	Clustered	Clustered	Clustered
N	332	332	332	324
Adj. R^2	0.992	0.992	0.984	0.868

Note: Standard errors in parentheses, clustering on province level; Dependent variables are log number of vehicles in 10000; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; City and year fixed effects are included; Controls include climate variables, demographic and macroeconomic variables.

Table 3.G.4: Difference-In-Difference (DID) Regression Results, Extension on New Registered Civil Use Automobiles

	(1) Total	(2) Passenger	(3) Cargo	(4) Other
Policy	-0.065	-0.172	0.024	0.224
	(0.105)	(0.111)	(0.080)	(0.181)
Control Provinces	Full	Full	Full	Full
Controls	Yes	Yes	Yes	Yes
SE	Clustered	Clustered	Clustered	Clustered
N	331	332	332	265
Adj. R^2	0.894	0.929	0.910	0.597

Note: Standard errors in parentheses, clustering on province level; Dependent variables are log number of yearly new registered civil use vehicles in 10000; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; City and year fixed effects are included; Controls include climate variables, demographic and macroeconomic variables.

Table 3.G.5: Difference-In-Difference (DID) Regression Results, Extension on Traffic Variables and Private Owned Automobiles

	(1)	(2)	(3)	(4)	(5)	(6)
	Number of Buses	Total Passengers Travel by Bus (10,000)	Number of Taxis	Total Private Vehicles	Total Private Passenger Vehicles	Total Private Cargo Vehicles
Chengdu	0.185***	0.169***	0.147***	-0.179***	-0.237***	-0.116
	(0.037)	(0.049)	(0.030)	(0.045)	(0.042)	(0.085)
Guiyang	-0.188***	-0.199***	0.653***	0.011	0.105	0.071
	(0.025)	(0.040)	(0.018)	(0.108)	(0.105)	(0.155)
Hangzhou	-0.086***	-0.029	0.051**	-0.053	-0.047	-0.318***
	(0.027)	(0.037)	(0.020)	(0.043)	(0.044)	(0.079)
Harbin	-0.091***	-0.151***	0.096***	-0.060	-0.105**	-0.028
	(0.023)	(0.039)	(0.019)	(0.048)	(0.051)	(0.087)
Tianjin	-0.030	0.076	-0.229***	-0.194***	-0.240***	-0.050
	(0.040)	(0.065)	(0.046)	(0.050)	(0.046)	(0.100)

Note: Standard errors in parentheses, clustering on city or province level; City column indicates treatment city for running individual DID for traffic variables, or the province they are capital of for auto market variables; Dependent variables are log number of vehicles or passengers in 10000; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; City and year fixed effects are included; Controls include climate variables, demographic and macroeconomic variables.

Table 3.G.6: Difference-In-Difference (DID) Regression Results, Extension on Individual Labor Time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Policy	-0.111**	-0.302***	-0.071*	-0.468***	-0.077	-0.393**	-0.049	-0.263***
	(0.048)	(0.078)	(0.040)	(0.171)	(0.066)	(0.176)	(0.062)	(0.081)
	[-0.207,-0.016]	[-0.455,-0.149]	[-0.150,0.008]	[-0.805,-0.131]	[-0.207,0.052]	[-0.741,-0.046]	[-0.172,0.073]	[-0.424,-0.102]
Sample	Full	Full	Urban	Urban	Rural	Rural	Narrow Window 2007-2012	Narrow Window 2007-2012
Controls	No	Yes	No	Yes	No	Yes	No	Yes
FE	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date
N	20432	6033	7300	2405	13130	3619	10847	3141
Adj. R^2	0.103	0.323	0.079	0.149	0.145	0.358	0.127	0.378

Note: Standard errors in parentheses, clustering on city-by-year level; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is log of last week labor hours; Controls include climate variables (daily average temperature and daily precipitation), and individual characteristics from the CHNS surveys.

3.H Robustness Checks

Table 3.H.1: Difference-In-Difference (DID) Regression Results Dropping Suspicious Manipulation Cities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Policy	-0.976	0.393	-4.666***	-0.852	-6.135***	-1.531	-6.117***	-0.478
	(2.532)	(3.093)	(1.791)	(2.585)	(1.756)	(2.759)	(1.712)	(3.084)
	[-6.008,4.056]	[-5.755,6.541]	[-8.200,-1.132]	[-5.951,4.247]	[-9.591,-2.679]	[-6.960,3.898]	[-9.478,-2.757]	[-6.533,5.576]
Control Cities	Only Policy	Only Policy	Same Province	Same Province	Matched	Matched	Full	Full
Controls	No	Yes	No	Yes	No	Yes	No	Yes
FE	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date
N	31637	31637	66423	66423	112702	112702	251652	251652
Adj. R^2	0.281	0.377	0.313	0.421	0.326	0.432	0.342	0.460

Note: Standard errors in parentheses, clustering on city-by-year level; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is daily average AQI; Controls include climate variables (daily average temperature and daily precipitation), and indicators for major pollutants.

Table 3.H.2: Difference-In-Difference (DID) Regression Results in Narrow Window 2007-2012

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
policy	1.862	1.338	-0.300	0.111	-0.585	0.293	-0.054	1.396
	(1.920)	(1.877)	(1.519)	(1.799)	(1.493)	(1.731)	(1.439)	(2.346)
	[-2.015,5.738]	[-2.454,5.129]	[-3.307,2.707]	[-3.452,3.674]	[-3.528,2.358]	[-3.120,3.707]	[-2.880,2.773]	[-3.214,6.006]
Control Cities	Only Policy	Only Policy	Same Province	Same Province	Matched	Matched	Full	Full
Controls	No	Yes	No	Yes	No	Yes	No	Yes
FE	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date	City, Date	City-by-Year , Date
N	15168	15168	43358	43358	71572	71572	182108	182108
Adj. R^2	0.249	0.389	0.314	0.457	0.321	0.457	0.337	0.462

Note: Standard errors in parentheses, clustering on city-by-year level; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is daily average AQI; Controls include climate variables (daily average temperature and daily precipitation), and indicators for major pollutants.

Table 3.H.3: Difference-In-Difference (DID) Regression Results, Robustness Check with US Embassy Hourly *PM2.5*

	(1)	(2)	(3)	(4)
Policy × Peak Hour	-7.299	-7.303	-7.016	-6.956
	(4.652)	(4.658)	(3.790)	(3.748)
	[-20.215,5.617]	[-20.238,5.631]	[-17.538,3.506]	[-17.361,3.448]
FE	City	City-by-Year	City	City
	Date	Date	Date	Date
	Hour	Hour	Hour	Hour
SE	Cl. City	Cl. City	Cl. City	Cl. City
Controls	No	No	Yes	Yes
Sampe	Full	Full	Full	Trim *PM2.5* > 900
N	167354	167354	161986	161982
Adj. R^2	0.421	0.425	0.429	0.431

Note: Standard errors in parentheses; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is hourly PM2.5 in $\mu g/m^3$.

Table 3.H.4: Difference-In-Difference (DID) Regression Results, Robustness Check with Annual *PM10*

	(1)	(2)	(3)	(4)
Policy	-1.433	-10.436*	-7.024	-9.739*
	(4.106)	(6.190)	(4.778)	(5.322)
	[-9.668,6.803]	[-22.745,1.872]	[-16.621,2.573]	[-20.216,0.739]
Control Cities	Only Policy	Same Province	Matched	Full
Controls	Yes	Yes	Yes	Yes
SE	Robust	Clustered	Clustered	Clustered
N	86	521	375	1715
adj. R^2	0.831	0.633	0.683	0.685

Note: Standard errors in parentheses; Clustering on city level; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is yearly average PM10 in $\mu g/m^3$; City and day fixed effects are included; Controls include climate variables and 3-year prior policy average of demographic and macroeconomic variables.

Table 3.H.5: Difference-In-Difference (DID) Regression Results, Robustness Check with Monthly AOD

	(1)	(2)	(3)	(4)
Policy	-0.063	-0.028	-0.041	-0.022
	(0.056)	(0.043)	(0.044)	(0.041)
	[-0.175,0.048]	[-0.113,0.057]	[-0.128,0.046]	[-0.103,0.059]
Control Cities	Only Policy	Same Province	Matched	Full
Controls	Yes	Yes	Yes	Yes
SE	Clustered	Clustered	Clustered	Clustered
N	977	10802	5711	37621
Adj. R^2	0.555	0.653	0.605	0.661

Note: Standard errors in parentheses; Clustering on city-year level; 95% confidence interval in squared brackets; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable is monthly average AOD on scale $[0, 1]$; City and month fixed effects are included; Controls include climate variables and city-by-year FE.