Title
Essays in Climate Economics

Permalink
https://escholarship.org/uc/item/0pj4n2gg

Author
Yu, Chu

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays in Climate Economics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Economics

by

Chu Yu

Committee in charge:

   Dr. Richard T. Carson, Chair
   Dr. Judson Boomhower
   Dr. Jennifer Burney
   Dr. Alexander Gershunov
   Dr. Joshua Graff-Zivin

2020

The dissertation of Chu Yu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

To my parents.

与我的父母分享

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

VITA

2014   B. A. in Economics and Mathematics, Northwestern University

2020   Ph. D. in Economics, University of California San Diego

ABSTRACT OF THE DISSERTATION

Essays in Climate Economics

by

Chu Yu

Doctor of Philosophy in Economics

University of California San Diego, 2020

Dr. Richard T. Carson, Chair

Chapter 1 of this dissertation seeks to understand the sources and nature of measurement errors in temperature variables, and how they can influence climate impact estimations. Measurement error in temperature variables is usually assumed to be "classical" i.i.d. normal and "small". This type of measurement error leads to attenuation bias, that is often small enough to be ignored. We show, however, that measurement errors in temperature variables are often large, distinctly non-normal, and vary systematically across space and time. The divergence between our empirical results and conventional wisdom stems from the fact that the construction of the temperature variables involves a series of steps, each of which introduces a distinct source of measurement error. This work is the first to formally characterize sources of measurement error

in temperature variables and how they interact. Simulation results are provided that illustrate the influence of these sources of measurement errors on climate impact estimates. We further propose a correction method that can be used to obtain consistent estimates of the parameters of interest under the conditions identified.

In Chapter 2, I zoom in to examine one particular source of measurement error: measurement error induced by using coarse data to approximate daily mean temperature. Much of the historical records from weather stations around the world reports on minimum and maximum temperature and this practice is still followed by most stations. The minimum and maximum are averaged to approximate daily mean temperature, the exposure variable most commonly used in climate impact studies. Empirically, I find substantive differences between daily mean temperatures constructed using hourly temperature data and daily minimum and maximum temperature. This single source of measurement error can easily result in a 5 to 10% bias in estimates of the impact of daily mean temperature on an output measure of economic interest, with the direction of the bias dependent on the location and season of the year.

Chapter 3 investigates climate impact models from another perspective: the representation of climate variables. The "bin" regression model has been put forward as a flexible semi-parametric method for representing a climate variable and it has emerged as the workhorse approach for empirical work (e.g., Deschênes and Greenstone, 2011). Our work is the first to formally explore econometric properties of the bin regression approach. We show that, although the bin regression approach often produces reasonable results, that the approach produces consistent parameter estimates only under very stringent and highly unlikely assumptions about the true data generating procedure. Problems with the bin regression approach are likely to be most severe in the tail bin categories, where most policy interest with respect to climate change impacts lies. We propose alternatives to bin model for the climate change impacts that produce consistent estimates and generally have better efficiency properties.

Chapter 1

Sources and Nature of Measurement Error in Estimating Climate Related Impacts

1.1   Introduction

Economic agents (e.g. farms, firms or individuals) respond in various ways to climate. There are thousands of studies that look at impact of weather variables (e.g. temperature, precipitation) and prospective climate change. Such papers appear in many fields, including agriculture, disaster management, ecology, epidemiology, health, recreation and transportation. In recent years, a large number of economics research has been undertaken to study the influence of climate variables on a wide range of different social and economic outcomes. These papers reshape our understanding of climate and human interaction, and provide useful empirical evidence on relevant climate policy. On agricultural impact, [Schlenker and Roberts, 2009] show that extreme heat hours causes damage to crop yields. On health impact, [Deschênes and Greenstone, 2011] examine climate and mortality relationship in the US, and show that both extreme cold and heat days cause extra mortality. On long run adaptation effort to climate change, [Barreca et al., 2016] provide the first large-scale evidence on how long-run adaptation alleviates hot day's impact on

health, and find that US's diffusion of residential air conditioning explains essentially the entire decline in hot day-related fatalities since 1960.

As noted in [Hsiang, 2016], a key methodological step in identifying temperature effect is the measurement of climate variables, regardless of the research design being used. An often documented, but rarely dealt with, problem is the deviation between the actual value of the climate variable of interest and the proxy variable used in the estimated climate impact models ([Auffhammer et al., 2013]). This deviation is known as measurement error in weather variables. Consider a linear model relating an outcome variable to temperature:

$$Y_{id} = \beta T_{id}^* + \varepsilon_{id}$$

where Y_{id} is observed for location i on day d. However, for various reasons which will be discussed later in the chapter, one only observe a temperature proxy T_{id} that introduces an additive measurement error u_{id} to the true temperature:

$$T_{id} = T_{id}^* + u_{id} \tag{1.1}$$

When estimating the impact parameter β using the mismeasured temperature, the estimated coefficient $\hat{\beta}$ is generally biased. A conventional wisdom is that measurement errors in climate variables are typically assumed to be classical measurement errors. Under the classical measurement error assumptions, the estimated coefficient $\hat{\beta}$ converges in probability to the true parameter β multiplied by an attenuation factor. This attenuation factor is a function of the variance of the true temperature $\sigma_{T^*}^2$ and the variance of the measurement error σ_u^2.

$$\hat{\beta} \xrightarrow{p} \beta \times \underbrace{\frac{\sigma_{T^*}^2}{\sigma_{T^*}^2 + \sigma_u^2}}_{\text{Attenuation Factor}} \tag{1.2}$$

2

As shown in equation (1.2), if measurement errors are classical, they attenuate the estimated impact coefficient towards zero since the attenuation factor is bounded by 0 and 1. The attenuation bias has generally been ignored in the existing literature, as it merely suggests that the estimated impacts identified from these climate impact models are conservative. [Lobell, 2013] conducts a simulation exercise on measurement error in temperature and precipitation variables, and finds adding small[1] and classical measurement error in temperature only slightly affects the coefficient measuring crop yield response to temperature. This finding is consistent with classical measurement error framework in which the size of attenuation bias only depends on the variance of the measurement error. However, the conditions underlying classical measurement error assumption are highly restrictive and are rarely tested in the climate economics literature. If measurement errors are not classical in nature, the size and sign of bias introduced by the presense of measurement error are no longer clear.

In this paper, we show that the common assumptions about measurement error do not hold empirically. Measurement errors in temperature statistics are often large in magnitude, distinctly non-normal, and vary systematically in predictable ways across time and space. The divergence between our finding and conventional wisdom occurs because measurement error comes in various different sources as researchers try to construct temperature statistics that can match with their research designs. For example, outcome variables of interest are measured at various temporal and spatial resolution, which are rarely matched well with climate data. As a result, researchers often need to assemble temperature data in certain ways to match with outcome variables and to identify climate impact at a level pre-determined by the level of outcome variables. When the spatial and temporal resolution matters in a particular study, the practice of assembling temperature via various steps would automatically introduce multiple sources of measurement error, which will consistently biased the impact estimation in uncertain directions

---

[1] The variance of measurement error in [Lobell, 2013] is $0.04°C$.

and magnitude.

This paper is the first in the climate economics literature to formally document sources of errors introduced in the construction of weather variables. We focus on temperature, a key climate change variable, although other climate variables might bear similar issues. Measurement error[2] in temperature variables is usually found in four different sources, following a logical order:

A **Monitor interpolation error**. Researchers often need to construct temperature variables with some spatial interpolation methods, because weather monitors are sparsely located. Therefore, the difference between the actual exposed temperature and the predicted temperature is taken as the monitor interpolation error.

B **Summary statistics estimation with coarse data**. Suppose the true stimulus variable is daily mean temperature[3] , which is always approximated by averaging minimum and maximum temperature in the literature. However, some recent meteorological studies ([Bernhardt et al., 2018]) have noted a statistically significant difference among different averaging methods, and estimating mean with two extreme order statistics would almost always introduce nontrivial measurement error.

C **Spatial aggregation error**. Spatial aggregation error is introduced due to a lack of high resolution outcome variable data. When outcome variables are available at county/state level, researchers will need to aggregate temperature over space to match with the outcome variables, thus measurement error is introduced in the aggregation process.

D **Error introduced from binning temperature**. The "bin" regression model has emerged as the workhorse approach in the climate economics literature (e.g., [Deschênes and Greenstone, 2011]). Binning is essentially a practice to construct temperature statistics

---

[2]In this paper, we deal with any types of error that can be mathematically represented as in equation (1.2). We acknowledge that some types of errors discussed here are misspecification error in nature.

[3]This differs from the first type of error in which the true stimulus variable is unknown.

which involves two steps: First, it puts daily temperature observations into dummy variables or pre-defined temperature bins; then it counts the number of days that fall into each temperature bin. Since measurement errors from other sources are likely introduced before binning, throwing mismeasured temperature observations into temperature bins would often lead to missclassification error with profound impact.

There are three focuses of this paper: (1) we formally characterize sources of measurement error and empirically test on the classical measurement error assumptions; (2) we explore how measurement error affects parameter estimates in climate impact studies and (3) we propose a correction factor method that can correct for both classical and non-classical measurement error.

To study the error properties of temperature statistics, we quantify several sources of measurement errors in temperature statistics by taking the difference between predicted temperature and the "true" temperature at station-level. We provide estimates of measurement error for more than 1,500 weather stations spanning 58 years in the contiguous U.S. We then conduct a series of tests to examine if classical measurement error assumption is valid for these empirically calculated measurement errors. We reject the null hypothesis that measurement errors in temperature variable are classical in every dimension, and find instead that these errors are often large, distinctly non-normal, and vary systematically across time and space.

We then study measurement error's effect on climate impact estimation when multiple sources of non-classical measurement error are present. There are two questions that are important to understand, and we answer these questions with two exercises. The first question is if measurement error in temperature variables leads to substantial change in the findings of existing climate impact literature. To this end, we implement an application under the U.S. agriculture context, in which we assemble several proxies of temperature following the literature convention and study if crop yield's response to temperature are sensitive to the choice of temperature proxy

in the regression. We find under a simple linear framework that some types of crops are extremely sensitive to how temperature is constructed. Using different temperature proxies, the difference in estimated temperature impact is 6.8 lb per acre (26% of the point estimates) for rice; 1.9 lb per acre (32% of the point estimates) for beans; and 0.14 lb per acre (33% of the point estimates) for cotton. The impact estimations are likely going to be more sensitive to the choice of proxy when more complicated models (e.g., adding covariates; using bin regression model) are used. Overall, this application confirms that the measurement errors are large in magnitude to have significant effect on the identification of temperature related impacts.

The second question is how the interaction across various sources of measurement errors affect the identification of climate impacts models. For example, some may think that aggregation over larger spatial or temporal scale may alleviate measurement error issue found in the more granulated temperature. To answer this question, we run a numerical exercise with actual temperature data and add sources of measurement error one at a time in a logical order. In general, we find that all four sources of measurement errors documented in this work can accentuate measurement error bias. The cumulative measurement error is about 30% of the natural variation in true temperature, and this cumulative measurement error can lead to about 10 to 40% of bias in either direction in the estimation of temperature's impact. In particular, the two most common types of measurement error—interpolation and binning, each contributes to about 10% of noise to the total variation in true temperature.

Finally, we propose an analytical framework to correct for the bias due to non-classical measurement error and use temperature data in Southern California as an example to illustrate our method. There are two common approaches in the applied literature to address measurement error in the independent variables: instrumental variable approach and correction factor approach. As we will show later in the paper, a valid instrument for temperature is not likely available for

non-classical measurement error so that the standard IV approach breaks down in this context. Therefore, we propose a correction factor method which uses consistent estimates of the second moments of measurement errors to correct for non-classical measurement error. This method is widely used in industrial production processes where the error statistics of the mismeasured variable can be readily calibrated.

The correction factor approach is simple in nature. Take the example of classical measurement error model, as illustrated in equation (1.2), if we can find consistent estimates of σ_{T*}^2 and σ_u^2, we can back out the true parameter β from the OLS estimate $\hat{\beta}$ by substituting in these variance terms into the fraction. The case with non-classical measurement error model is similar except that a few more terms need to be estimated to construct the correction factor, which we characterize in detail in section 5. To consistently estimate these variance and covariance terms, we construct a Multivariate Adaptive Regression Splines (MARS) model to predict each of these terms as a function of the method used to construct temperature statistics, latitude, longitude, and other factors that may influence the magnitude of measurement errors. We then use the correction factor method to correct for measurement error bias with a cross validation exercise using temperature data from 258 monitors in Southern California. Our method on average can reduce measurement error bias by 30% in this exercise, and its performance can likely been improved by a richer prediction model for correction factor.

This work addresses important methodological issues that can be useful for broader audience. First, it is relevant for any problems where spatial and temporal resolution matter but that measurements of some variables are coarser than desired. Economic agents (e.g., farms, firms or individuals) respond in various ways to exogenous stimulus. The problem of matching agents with the correct stimulus variable often occurs to empirical researcher. For example, in some large survey data, individuals are recorded at various spatial resolutions (e.g., census track, zipcode,

city or county), but are actually exposed to true stimulus variables that cannot be measured exactly. This divergence between the observed exposure and true exposure introduces a type of measurement error which can have profound impact on the estimation of relevant parameters. In this work, we show empirically the potential bias introduced as a result of the divergence in spatial resolution across variables under the climate economics context. Further, we devise a procedure to correct for the measurement error bias due to spatial mismatch, which is applicable to any studies that involves estimating the impact of stimulus variables (e.g., advertising, contagious disease, pollution, weather).

Second, this work is also relevant for studies that include weather variables as control variables. Classic textbooks on measurement error (e.g. [Fuller, 1987], [Carroll et al., 1995]) have well established that in multivariate linear regression with error, biases can be introduced in the coefficients of both the error-prone and error-free variables, and these biases can be additive or multiplicative depending on the relationship among the covariates. Although measurement error issue in multivariate regression is more complicated to deal with and beyond the scope of this paper, we provide summary statistics on measurement error which would help pick the "best" procedures to assemble temperature proxy that incur small measurement error bias.

Third, one of this work's end-product is a modelling system that can produce an estimate of the first and second moments of measurement error for temperature for a particular interpolation method at any specified U.S. location and date. Applied researchers can use these estimates to construct correction factors for use in their estimates of temperature impacts. The methodology developed should be extensible to other weather and pollution variables.

The paper is organized as follows. Section 1.2 is a literature review of measurement error issue in climate economics. Section 3 characterizes measurement error in temperature variables,

and show that there are various sources of non-classical measurement errors. In section 1.4, we use two numerical exercises to show how measurement error as characterized in section 1.3 affects the estimation of climate impact models. In section 1.5, we propose a correction factor framework that can correct for measurement error in the linear climate impact models. Section 1.6 concludes this chapter.

1.2   Literature Review

Measurement error is one of the most fundamental problems in empirical economics. Presense of measurement error causes biased and inconsistent parameter estimation, which may lead to incorrect policy implications to various degrees. Because of its importance, measurement error problems have been well documented in econometrics and statistics literature. Major early studies on measurement error can be found in [Griliches, 1974], [Fuller, 1987], [Carroll et al., 1995] and [Hausman, 2001] to name only a few. In terms of recent development in theoretical work, [Schennach, 2016] reviews recent progress made in developing estimation methods for nonlinear models in the presense of measurement errors.

Issues of measurement error have been well documented in some empirical fields, such as labor economics, macroeconomics and industrial organization. Early work on the measurement error issue in the applied field has focused on the typical textbook model of classical measurement error. However, in [Bound et al., 1994]'s study on the error properties of labor market data, their results show a need to recognize the potential importance of measurement error and to incorporate more realistic assumptions about the properties of such error into measurement error models. [Bound et al., 2001] further concludes that the possibility of non-classical measurement error should be taken more seriously, both in assessing the likely biases in analyses that take

9

no account of measurement error and in devising procedures that "correct" for such error. In macroeconomics, [Ahn and Hamilton, 2019] catalog inconsistencies in U.S. unemployment rate, labor-force participation rate and duration of unemployment data, and propose a unified reconciliation. [Hu, 2017] describes a few applications of nonparametric identification results involving latent variables, such as belief, productivity and unobserved heterogeneity in empirical industrial organization and labor economics.

In climate economics literature, presense of measurement error in climate variables is usually ignored. Existing literature tends to assume that measurement error is classical, which leads to well known econometric consequence — attenuation bias. In assuming classical measurement error, if one finds nonzero climate effect, such effect is usually confirmed because attenuation bias leads to an under-rejection of the null hypothesis, and the sign of such effect is usually preserved. As a result, in climate economics literature, attenuation bias is small enough to be ignored after noting that the parameter estimates of interest are likely to be conservative. [Lobell, 2013] uses a simulation extrapolation (SIMEX) approach to gauge the importance of measurement error for two recent studies that employed statistical crop models. The study finds that current levels of measurement error are unlikely to bias predictions related to temperature changes, but more likely to influence predictions to precipitation changes. Although the study is still within the realm of classical measurement error assumption, the main result does suggest that greater attention to measurement error is waranted in climate economics.

1.3   Characterization of Measurement Error

In this section, we formally characterize properties of measurement error in temperature variables. Section 1.3.1 is a complete description of four sources of measurement error. Section 1.3.2 describes the procedure for constructing measurement error empirically. Section 1.3.3 is a

detailed discussion on the classical measurement error assumption and tests conducted on each classical measurement error assumption.

1.3.1 Sources and Types of Measurement Errors

The construction of temperature proxy variables almost always involves a series of steps, each of which introduces a different source of measurement error. This section discusses sources of measurement error introduced in constructing temperature proxy variables following a logical order. As noted in the introduction section, we take a mathematical definition of "measurement error" as described in equation (1.1). In other words, the difference between the exposed temperature and the constructed temperature is defined as "measurement error" in this work. Note that we fully acknowledge that some types of errors discussed here are misspecification error in nature, and the use of the word "measurement error" is purely semantic in some types of errors.

Monitor interpolation error

Monitor interpolation error is a type of measurement error due to the sparsity of ground weather stations. It is introduced because there is almost never a weather monitor at exactly where researchers want. If there is only one weather monitor within each study region, researchers would simply assign the reading of that monitor to the entire study region. When there are many monitors or no monitor around the study region, they use some interpolation techniques to construct temperature and other weather variables.

The most commonly used interpolation methods in climate economics literature include closest station, simple averaging, and inverse distance weighted (IDW) averaging. To implement any interpolation method, one needs to define a point of interest (some centroid), and draw a circle with arbitrary radius (often 100km or 200km) as boundary. Any weather stations that fall into the circle can be used for interpolation.

11
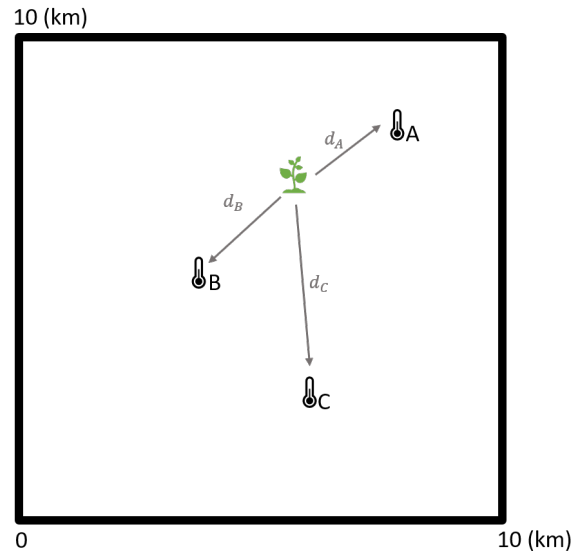
Figure 1.1: Monitor Interpolation Error $\hat{T} - T^*$

We use Figure 1.1 to illustrate some common interpolation methods. Suppose there are three monitors labeled A, B and C in a 10×10km block. The goal is to predict temperature at the farm (the icon in green). There are various ways to interpolate temperature at this farm:

- **Closest Station**. This method is the most widely used interpolation method because of its simplicity. Closest station method takes the reading of the closest station as the temperature for point of interest:

$$\hat{T} = T_A^*$$

- **Simple Averaging**. Simple averaging takes the average temperature readings of all stations that fall into the block:

$$\hat{T} = \frac{T_A^* + T_B^* + T_C^*}{3}$$

- **Inverse Distance Weighted (IDW) Averaging**. IDW averaging is a weighted average of

station readings, and stations closer to the centroid are given more weight:

$$\hat{T} = \frac{\frac{T_A^*}{d_A^p} + \frac{T_B^*}{d_B^p} + \frac{T_C^*}{d_C^p}}{\frac{1}{d_A^p} + \frac{1}{d_B^p} + \frac{1}{d_C^p}}$$

d_i is the distance between farm and station i for $i = A, B, C$. p is a power parameter that determines how much we would like to penalize the distance. Empirical studies typically choose p=2.

- **Model Based Spatial Smoothing**. Over the past decade, there have been large technological improvements in the measurement of climate variables. Spatial smoothing models bring a combination of climatological and statistical concepts to the prediction of weather variables. An often used weather product is PRISM[4], which stands for Parameter-elevation Regressions on Independent Slopes Model. It is widely regarded as one of the best geographic interpolation procedures. The PRISM Climate Group gathers climate observations from a wide range of monitoring networks, applies sophisticated quality control measures, and develops a comprehensive spatial climate datasets for the contiguous U.S. with a fine resolution of 4×4 km. PRISM data are widely used in recent climate economics literature, mainly because it is a gridded weather data product that are easily accessible for economists. In our example, to estimate temperature at the centroid using gridded data product like PRISM, one needs to find the corresponding grid on PRISM that covers the centroid, and takes the reading of that grid as the temperature of the centroid.

The first three methods are linear combinations of temperature readings of existing stations. Closest station method places all weight on the closest station, and zero to the rest. Simple averaging weights all stations equally. IDW averaging has weight in between the first two methods. PRISM incorporates information from not only adjacent monitors, but also local climate and GIS variables. In section 1.3.2, we construct measurement error introduced by these

---

[4]PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, created 10 Oct 2018

different interpolation methods using contiguous U.S. weather station data, and will show that PRISM tends to lead to smaller measurement error bias than other interpolation methods.

Summary Statistics Estimation with Coarse Data

This is a type of measurement error due to a lack of high frequency temperature data. A most oftenly used summary statistics of temperature is the daily mean temperature. As noted in some recent meteorological studies ([Bernhardt et al., 2018]), there is a statistically significant difference across different averaging methods. Suppose daily mean temperature is the correct stimulus variable under true DGP, which should be computed as:

$$T_{mean} = \frac{1}{S} \int_0^S f(s) ds$$

where S is the number of seconds in a day. The true mean should be an integral over the temperature in each second. However, because many weather stations and PRISM only report daily minimum and maximum temperature, daily mean temperature has to be computed with a much coarser dataset:

$$\hat{T}_{mean} = \frac{\tau_{min} + \tau_{max}}{2}$$

$\hat{T}_{mean}$ is an unbiased estimator only if τ_{min} and τ_{max} are symmetrically distributed. Apparently, this condition is strange and highly restrictive. In broader statistical sense, to estimate any mean with two extreme order statistics is almost never a good approach.

[Ma and Guttorp, 2013] show that daily temperature curves are not symmetric about the average daily temperature. The study takes daily minute average temperature as ground truth, and compare different temperature averaging methods. They find that using hourly temperature as proxy introduces the smallest bias, while using only τ_{min} and τ_{max} leads to the largest bias. Their

result suggests that using hourly temperature should be adequate for representing the true daily temperature, but not average of τ_{min} and τ_{max}.

Spatial Aggregation Error

This is a type of measurement error that is due to a lack of high resolution outcome variable data. In practice, outcome variables are often aggregated at city/county/state level. Therefore, rather than assigning a temperature value to a point, researchers often need to assign a temperature value to a large area like city/county/state. For example, estimating San Diego county's temperature and mortality relationship introduces aggregation error due to the fact that there is no monitor that measures temperature at "San Diego county"; all that people can observe is temperature readings from some sparsely located monitors in San Diego county.

To estimate temperature at any given county, it is a common practice to specify county centroid, and then interpolate temperature for that county centroid. Nevertheless, there are three potential problems with such practice: First, there are usually several county centroids to choose from, such as population centroid, geographical centroid and farm centroid. It is not clear if the chosen county centroid is truly the county centroid for the outcome variables of interest; second, any attempt to estimate temperature readings for a pair of spatial coordinate (county centroid in this case) would introduce interpolation error, unless there is a monitor exactly at the county centroid. Third, the level of aggregation varies across different locations. For example, while temperatures aggregated from counties in Rhode Island might only suffer small aggregation error since counties are rather small in terms of land area in Rhodes Island, temperatures estimated from counties in California usually bear much larger aggregation error.

There is a debate on whether the presence of spatial aggregation error increases or de-

creases measurement error bias. Some may think that small scale measurement error cancel each other when temperature is aggregated at a larger spatial scale. In other words, it can be hard to interpolate temperature at a point accurately, but it might be relatively easy to estimate daily average temperature for the U.S. as a whole. On the other hand, as we have discussed earlier, different choices of aggregation procedure often produce different proxies of temperature. Unless aggregation is conducted with great caution, aggregation error would exacerbate measurement error issue.

Error Introduced from Binning Temperature

Binned regression is the current workhorse model in the climate economics literature when outcome variable is observed less frequently than weather variables. A typical setting for binned regression is the following:

For location $i = 1, 2, \ldots, N$, year $t = 1, 2, \ldots, T$, observe a scalar outcome Y_{it}; In addition, for each i and t, observe T_{id} at daily frequency $d = 1, 2, \ldots, 365$. Choose summary statistics $\text{TBIN}_{j,it}$ as regressors in a linear model:

$$Y_{it} = \sum_{j=1}^{J} \beta_j \text{TBIN}_{j,it} + \eta_{it}$$

Here $\text{TBIN}_{j,it}$ denotes the number of days in jth bin at location i, year t.

Typical binned regression involves three steps, all of which introduces measurement or misspecification error to some extent. The first step is to put each daily weather observation into arbitrarily defined temperature bins. The second step is to count number of observations in each temperature bin for every location i and year t. The third step is to estimate outcome variable as a linear function of these temperature bins.

If we take one step back and think about the underlying assumption behind binned regression, it is justified only if outcome variable is a function of distribution of daily temperature. The second step is essentially summarizing a distribution of temperature for location i and year t. Yet variables $\text{TBIN}_{j,it}$ are merely one of the many possible summary statistics of the distribution, and in fact, it is quite unnatural to summarize any empirical distribution with piecewise constant function. We show this with a simple example:

Table 1.1 lists three series of temperature observations. The first row represents the baseline temperature scenario; second and third rows are two climate projections in which temperature increases in some days. In projection 1, temperature in day 1 increases by $9°$F. However, if we look at the binned distribution on Figure 1.2, this huge rise in temperature is not captured in the binned specification.

Under projection 2, temperature in day 2 increases by only $1°$F. Now that day 2 temperature hits the next bin threshold, the binned distribution changes drastically from baseline as shown in Figure 1.2, even though the actual increase in temperature is not as much as that in projection 1.

Table 1.1: Temperature Under Different Scenarios

	Day 1	Day 2	Day 3
Baseline	81	90	91
Projection 1	90	90	91
Projection 2	81	91	91
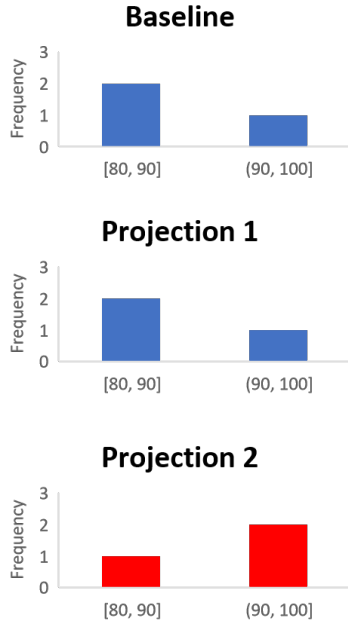
Figure 1.2: Histogram of Temperature

The third step is estimating outcome of interest as a linear function of these binned variables. It is at best a waste of degree of freedoms, since the more bins we include, the more parameters we need to estimate. In fact, estimating outcome variable as a linear function of probability distribution parameters often introduces large misspecification error.

This paper takes the perspective of measurement error, so we focus on the first step of binned regression. We take a simpler binning specification with step one only. Note that this setup does not collapse the dependent variable but rather is consistent with something very simple:

For location $i = 1, 2, \ldots, N$, day $d = 1, 2, \ldots, 365$, observe a scalar outcome Y_{id} and temperature T_{id} at daily frequency. Put T_{id} into J different bins $\text{TBIN}_{j,id}$:

$$Y_{id} = \sum_{j=1}^{J} \beta_j \text{TBIN}_{j,id} + \varepsilon_{id}$$

There are two layers of measurement error in this equation. First, by putting T_{id} into

TBIN$_{j,id}$, we are implicitly assuming T_{id} equals the mean or median of this temperature bin j. For instance, when we put $81°$F into temperature bin $80 - 90°$F, we have implicitly turned $81°$F into $85°$F. This potential loss of information matters when a cut-off temperature is between $81°$F and $85°$F so that the outcome's response changes significantly once the temperature is beyond the cut-off temperature.

Second, measurement error introduces misclassification under binned regression, which is large and asymmetric at end bins. Since there is no lower and upper bounds for the first and last bins respectively, temperatures with large measurement error are likely to concentrate in these two bins. As a result, erroneous policy conclusions can be drawn since end bins are the ones of interest from a climate change policy perspective.

1.3.2   Constructing Measurement Error Empirically

Data Sources

To construct measurement error empirically, we bring together several data sets on U.S weather:

Global Historical Climatology Network (GHCN). NOAA's Global Historical Climatology Network (GHCN) is an integrated database of climate summaries from land surface stations across the globe. We use GHCN Daily dataset for constructing monitor interpolation error. The variables of interest are daily minimum and maximum temperature. Due to limitation in computing power, we select a sample of ground weather stations to construct monitor interpolation error. We use population weighted random sampling procedure to draw a sample of the U.S. statistical regions. Office of Management and Budget (OMB) defines 382 Metropolitan Statistical Areas

and 535 Micropolitan Statistical Areas based on population and degree of social and economics integration. These 917 statistical areas are treated as the base sample for the random sampling procedure. We use 2010 census population as weight, and randomly sample 64 areas out of the 923 statistical areas.

We extract weather data from these 64 sampling units spanning the year 1960-2017. Our final sample includes 1516 stations that have valid readings and at least one valid nearby station. Figure 1.17 in section 1.7 is a map showing the stations we have included in the sample, with all 9 NOAA climate regions under the coverage.

Integrated Surface Database (ISD). NOAA's Integrated Surface Database (ISD) consists of global hourly and synoptic observations compiled from numerous sources. NOAA ISD-LITE is a simplified version of the ISD hourly dataset. We use 237 U.S. airport stations spanning the year 1981 to 2010. Weather records at airports are relatively accurate because these monitors are better maintained than monitors elsewhere. We use NOAA ISD-LITE hourly data to construct measurement error introduced by summary statistics estimation with coarse data. This dataset is also used for numerical exercise in section 4. Figure 2.1 in section 1.8 is a map showing the airport stations we have included in the sample.

Parameter-elevation Regressions on Independent Slopes Model (PRISM). The PRISM Climate Group gathers climate observations from a wide range of monitoring networks, applies sophisticated quality control measures, and develops spatially smoothed climate datasets. It has a fine resolution of 4×4 km grid across the contiguous U.S. We extract PRISM grids that match with ground weather station data for the year 1981 to 2017.

Computing Monitor Interpolation Error

We use a leave-one-out method to compute the predicted temperature for each ground weather station in the sample with four predictors, including closest station, simple averaging, IDW averaging, and closest elevation.
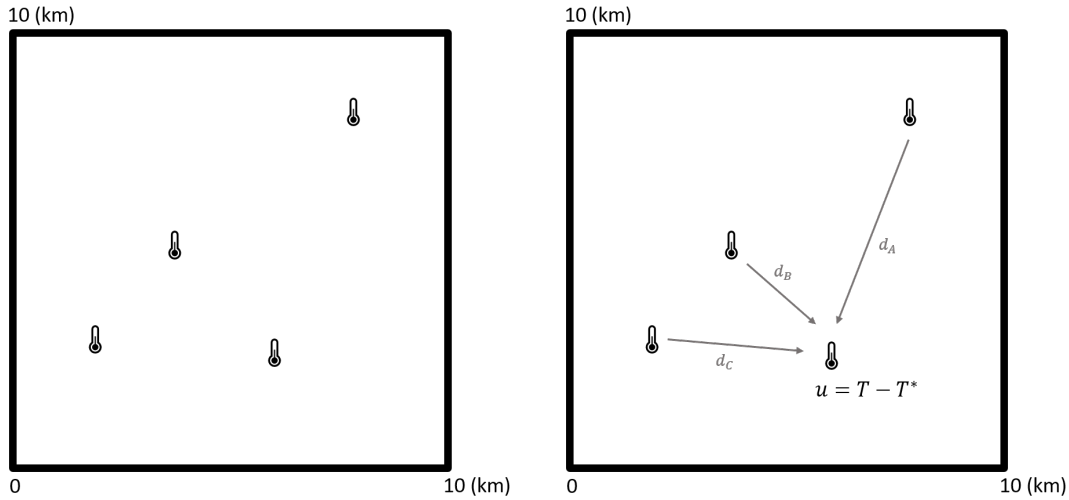
Figure 1.3: Computing Monitor Interpolation Error

To interpolate temperature at a target monitor, we draw a circle with a radius of 100km around that monitor. We use neighboring monitors' readings for prediction, and compare the predicted values to the actual readings at the target monitor. Figure 1.3 is a simplified illustration of this leave-one-out method. We take the bottom right monitor as the target monitor, drop this monitor, and use the remaining three monitors to predict the temperatures at the target monitor. For PRISM predictor, we extract predictions from PRISM grids that match with the target monitor. We quantify the interpolation measurement error as the difference between the target monitor's true reading and various predictions.

Measurement errors in minimum and maximum temperatures are calculated seperately. Table 1.2 is a summary statistics for constructed monitor interpolation errors in daily maximum

temperature data. Summary statistics for errors in daily minimum temperature are reported in section 1.8 Table 1.8. "T*" is the true temperature at monitors measured in celsius degrees and is taken as the ground truth temperature. "Closest Station","Simple Averaging", "IDW Averaging", "Closest Elevation" and "PRISM" denote absolute values of measurement error using these five methods. "N. Stations" is the number of stations that fall into the pre-defined circle, or in other words, the number of stations that can be used for interpolation for each target station. "Closest Dist" is the distance between the target station and its closest station. "Closest Elev" is the difference in elevation between the target station and station that bears the smallest elevation difference from the target station.

Table 1.2: Summary Statistics for Monitor Interpolation Errors (Tmax)

	Mean	SD	Min	Max	N.obs
T* (°C)	19.34	(10.82)	-33.30	53.90	15,243,148
Closest Station (°C)	2.22	(2.41)	0.00	34.10	15,243,148
Simple Averaging (°C)	2.11	(2.01)	0.00	27.52	15,243,148
IDW Averaging (°C)	1.87	(1.85)	0.00	31.93	15,243,148
Closest Elevation (°C)	2.60	(2.70)	0.00	36.70	15,243,148
PRISM(°C)	1.85	(2.27)	0.00	30.90	10,191,945
N.Stations	31.71	(17.79)	1.00	97.00	15,243,148
Closest Dist (km)	16.81	(10.18)	0.00	96.78	15,243,148
Closest Elev (feet)	3.50	(4.44)	1.00	127.00	15,243,148
Elevation (feet)	411.31	(616.73)	0.00	3507.90	15,243,148

- **Closest Station**. We take temperature readings of the closest station within the pre-defined boundary as the predicted temperatures. The smallest "closest distance" is obtained in downtown Philadelphia, where the two monitors differ only in elevation. The largest "closest distance" is obtained in Los Angelas county, where two stations are 96.7km apart. The average "closest" distance is about 17 km in the sample.

- **Simple Averaging**. This method involves taking the average of all weather stations within the pre-defined boundary. In constructing measurement error, we drop the stations that

do not have a nearby station within 100km. In other words, for each station, there is at least one other station that can be used for temperature prediction. The average number of stations used for this simple averaging method is 32, with a maximum of 97 adjacent stations used for prediction in Los Angeles county.

- **IDW Averaging**. IDW Averaging is similar to simple averaging, except that closer stations are given more weight, where weight is defined as the inverse of squared distance between target station and neighboring stations.

- **Closest Elevation**. We calculate the difference in elevation between target stations and neighboring stations, and take the reading of station that bears the smallest elevation difference from the target station as the fourth temperature predictor. Using closest elevation for temperature prediction is uncommon in climate economics literature, but this exercise is insightful as GIS information other than the distance between stations is embedded in this predictor. In our dataset, elevation ranges from 0 to 3508 feet. The station with the highest elevation is located at Clear Creek, CO, which is close to Denver. The smallest elevation difference is 0 feet, as is the case in most plains. The largest elevation difference is 127 feet and occurs in Washoe County, Nevada.

- **PRISM**. We match each ground weather monitor with its corresponding grid in PRISM dataset, and compute the measurement error introduced by PRISM. Overall, PRISM dataset bears the smallest value of absolute measurement error among the five interpolation methods.

Computing Coarse Data Mean Estimator Error

Mean temperature is the most oftenly used climate variable in the literature. Using coarse data to estimate daily mean temperature introduces a source of measurement error. Therefore, in this section, we compute measurement error caused by measuring mean temperature with coarse

data.

[Ma and Guttorp, 2013]'s main result suggests that using hourly temperature should be adequate for approximating the true daily temperature. Therefore, we take 24 hours daily mean temperature as the ground truth temperature:

$$T_{mean} = \frac{1}{24} \sum_{h=1}^{24} \tau_h$$

and compare that with estimation of mean by averaging minimum and maximum temperature:

$$\hat{T}_{mean} = \frac{\tau_{min} + \tau_{max}}{2}$$

Table 1.3 is a summary statistics for constructed measurement error using both airport weather station data and PRISM data. "Mean Temp" is the 24 hours daily mean temperature, which is defined here as the ground truth temperature. "Tmin" and "Tmax" are the lowest and highest temperature within 24 hours. $|u|$ denotes the absolute value of constructed measurement error, and elevation is measured in meters in this dataset.

Table 1.3: Summary Statistics for Coarse Data Mean Estimator Error

	Mean	SD	Min	Max
Mean Temp (°C)	12.50	(10.89)	-47.13	41.23
Elevation (m)	378.43	(472.02)	0.30	2296.10
Weather Station:				
Tmin (°C)	7.56	(10.71)	-50.60	36.10
Tmax (°C)	18.52	(11.56)	-43.80	50.00
$|u|$(°C)	0.77	(0.67)	0.00	22.86
PRISM:				
Tmin (°C)	7.62	(10.82)	-44.77	30.20
Tmax (°C)	18.77	(11.35)	-31.56	43.78
$|u|$(°C)	1.47	(1.35)	0.00	38.67
Observations	2,280,749			

The mean and standard deviation of maximum and minimum are close between airport station data and PRISM data. In terms of the magnitude of measurement error, PRISM bears larger measurement error than using weather station. Most likely because PRISM introduces two sources of measurement error: interpolation error and mean estimator error introduced by using coarse data.

Size of measurement error caused by coarse data mean estimators appears small compared to other sources of measurement error, however it varies across space and time depending on microclimate conditions. For example, the maximum measurement error generated with weather station occurs in September 1981 at Winslow, AZ. The 24 hours daily mean temperature is about 13°C. However, because lowest temperature drops to -45°C at some point, using coarse data mean estimator leads to a daily mean temperature of -10°C, inducing a measurement error as large as 22.86°C as reported in Table 1.3.

1.3.3 Classical Measurement Error

We start with the simplest climate impact estimation model in which temperature is the only independent variable. Suppose we wish to estimate coefficient β in the response function:

$$Y_{id} = \beta T_{id}^* + \varepsilon_{id} \tag{1.3}$$

However, we only observe:

$$T_{id} = T_{id}^* + u_{id} \tag{1.4}$$

In classical measurement error model, we assume:

A1 $u_{id} \sim i.i.d.N(0,\sigma^2),\ \forall i,d$

A2 $cov(T_{id}^*, u_{id}) = 0$

A3 $cov(\varepsilon_{id}, u_{id}) = 0$

A4 $E(u_{id}^2) < \infty$

A1 states that measurement errors are independently and identically distributed with normal mean zero and variance σ^2. We show later with sets of empirically constructed measurement errors that ithey do not follow mean zero, do not have homoscedastic variance, and do not follow normal distribution in general.

A2 states that true temperature and measurement error are not correlated. Our empircal work shows that measurement errors are highly correlated with GIS variables (longitude, latitude, and elevation), and with true temperature. In fact, both extremely high and low temperatures are more prone to measurement error.

A3 states that measurement errors are uncorrelated with idiosyncratic error in the response variable. This assumption has often been taken for granted, as attention to non-classical measurement error is usually given to A2. Yet A3 can not be overlooked under certain restrictive context. For example, let the outcome variable be some health outcome, and proxy temperature as regressor. Some locations far away from the population centroid are more likely to be measured with large measurement error, and such locations are also less likely to receive timely and adequate medical treatment. Therefore, measurement error is correlated with idiosyncratic error, which can also affect climate impact estimation.

A4 states that the variance of measurement error is finite, and this condition is considered valid throughout this paper.

Attenuation Bias under Classical Measurement Error Assumption

Substitute equation (1.4) into equation (1.3) yields:

$$Y_{id} = \alpha + \beta(T_{id} - u_{id}) + \varepsilon_{id} = \alpha + \beta T_{id} + e_{id}\ ^{5} \tag{1.5}$$

The new error term $e_{id} \equiv \varepsilon_{id} - \beta u_{id}$ introduces endogeneity bias in the estimation of β. Under classical measurement error assumption, $\hat{\beta}$ can be written explicitly as:

$$\hat{\beta} = \frac{cov(T_{id}, Y_{id})}{var(T_{id})} = \beta \times \frac{\sigma_{T*}^2}{\sigma_{T*}^2 + \sigma_u^2} \tag{1.6}$$

The quantity $\frac{\sigma_{T*}^2}{\sigma_{T*}^2 + \sigma_u^2}$ is referred to as reliability ratio or signal-to-total variance ratio. This term is always between 0 and 1, and generates attenuation bias. When some studies mention "due to measurement error, these estimates are relatively conservative", attenuation bias as specified in equation (1.6) is what they actually refer to.

When variation of measurement error is relatively small compared to variation of true temperature ($\sigma_u^2 << \sigma_{T*}^2$), the term $\frac{\sigma_{T*}^2}{\sigma_{T*}^2 + \sigma_u^2}$ is dominated by σ_{T*}^2, thus attenuation bias due to measurement error is minimal; when variation of measurement error is relatively large compared to variation of true temperature ($\sigma_u^2 >> \sigma_{T*}^2$), $\hat{\beta}_{OLS}$ shrinks towards zero.

Some climate impact studies would also add location and time fixed effect into their main specifications to control for omitted variable bias. A concern with such practice is that fixed effects would absorb temperature variation across time and location, so that σ_{T*}^2 becomes minimal. The attenuation factor $\frac{\sigma_{T*}^2}{\sigma_{T*}^2 + \sigma_u^2}$ is then dominated by variance of measurement error, which leads

---

[5]Even though true DGP in equation (1.3) does not include a constant term, a constant term b_0 is included to ensure $E(e_{id}) = 0$.

to large attenuation bias by equation (1.6).

We calculate attenuation factor under classical measurement error assumption using our constructed measurement error. In fact, as long as consistent estimator for σ_u^2 and $\sigma_{T^*}^2$ are available, we can construct correction factor by plugging these consistent estimators into equation (1.6) and correct for measurement error under classical measurement error assumption. In section 5, we extend this method for non-classical measurement error models as well.

We will show in section 1.3.3 that most of the classical measurement error assumptions are not valid. Yet manually calculating attenuation factors for any single source classical measurement error assumption could still provide some insight on the magnitude of bias introduced by various sources of measurement error. We calculate attenuation factor as characterized by equation (1.6).

Table 1.4: Empirical Attenuation Factor

	(1) att.1	(2) att.2
Closest Station	0.92	0.88
Simple Avg	0.93	0.93
IDW Avg	0.94	0.94
Closest Elevation	0.89	0.85
PRISM	0.93	0.91
(Tmin+Tmax)/2	0.99	0.99

Column (1) in Table 1.4 reports attenuation factors calculated with constructed measurement error based on equation (1.6). We find that coarse data mean estimator bears the smallest attenuation bias, with an attenuation factor of 0.99. PRISM has an attenuation factor of 0.96, which is the best among all interpolation methods. IDW averaging estimator has an attenuation factor of 0.94, and is the best among the more traditional monitor interpolation methods. It is

perhaps a little surprising that closest station estimator, which is the most popular estimator for its convenience, generates a large attenuation bias of almost 10%. We should keep in mind that these are the best case scenario estimation — classical and single source. We show in section 1.4 that the presense of multiple sources of nonclassical measurement error would incur larger bias.

Tests on Classical Measurement Error Assumptions

In this section, we describe test procedures and test results on classical measurement error assumptions, including mean zero, normality, homoscedasiticy, independence, and independence from latent variable assumptions. The only assumption we cannot test directly is assumption A3, which involves knowledge of idiosyncratic error term ε_{id}. In fact, one can test assumption A3 straightforwardly using a Pearson correlation test with outcome variable data.

We test five classical measurement error assumptions with our constructed measurement error. The test statistics is reported on Table 1.5. For interpolation methods, we present results using maximum temperature. Test statistics for errors in daily minimum temperature are reported in section 1.8 Table 1.9.

Table 1.5: Tests on Classical Measurement Error Assumptions (Tmax)

	$E(u) = 0$			Hetero.	Normality		Serial Correlation		$cov(u, T^*) = 0$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Mean(u)	SD(u)	p-value	p-value	V stats	p-value	Lag Corr	p-value	Corr(u,T*)	p-value
Interpolation										
Closest Station	0.02	3.27	0.00	0.00	90.96	0.00	-0.44	0.00	-0.15	0.00
Simple Avg	-0.01	2.92	0.00	0.00	116.60	0.00	-0.42	0.00	-0.25	0.00
IDW Avg	-0.00	2.63	0.02	0.00	175.56	0.00	-0.43	0.00	-0.23	0.00
Closest Elevation	0.04	3.75	0.00	0.00	137.09	0.00	-0.42	0.00	-0.17	0.00
PRISM	-0.16	2.92	0.00	0.00	193.21	0.00	-0.43	0.00	-0.14	0.00
Coarse data estimator										
(Tmin+Tmax)/2	0.54	0.86	0.00	0.00	79.94	0.00	-0.46	0.00	-0.08	0.00

1. $E(u_{id}) = 0$. We test mean zero assumption with a simple t-test on the sample mean of

measurement error for each set of measurement error. In Table 1.5, column (1) through (3) show relevant test statistics for the t-test. The p-value indicates that we can safely reject the null hypothesis that sample mean is zero at a significant level of 5% for all these sources of measurement error.

In terms of the mean of measurement error, traditional monitor interpolation error has means that are closer to zero than either coarse data mean estimator or PRISM reanalysis data. Within traditional monitor interpolation methods, IDW averaging has the smallest mean measurement error overall. Coarse data mean estimator, which in this case is the average of minimum and maximum temperature, bears the largest deviation from mean zero. PRISM's mean measurement error also has some moderate deviation from zero. One interesting observation is that some estimators are more likely to produce overestimation, such as closest station, closest elevation and coarse data estimator. Simple average and PRISM tend to underestimate true temperature.

However, as equation (1.6) implies, it is the variance of measurement error σ_u^2, rather than the level of measurement error, that determines the size of bias under classical measurement error assumption. We show in section 1.5 that in the more general cases where one or more classical measurement error assumptions are relaxed, it is also the second moment of measurement error that is relevant for bias correction. Therefore, the variance of measurement error σ_u^2 are more informative of the size of bias caused by measurement error than the mean.

Although measurement error from coarse data mean estimator has the largest deviation from mean zero, it has the smallest range of error distribution. The standard error of coarse data mean estimator's measurement error is less than 1°C. This is followed by PRISM data, which also bears a relatively small standard error of 2.16°C , most likely because it is a spatial smoothing estimator. Among the more traditional monitor interpolation methods, IDW averaging has the smallest standard error. Estimators that only use information from one, i.e. closest distance

and closest in elevation, bear relatively large standard error, which implies that such estimators generate larger measurement error bias.

2. $var(u_{id}) = \sigma^2$. It is often assumed that measurement error is homoscedastic, but our heterscedasticity test shows this is not the case. In fact, variance of measurement error can be correlated with time, space, and other GIS variables. To test homoscedasticity assumption, we detect a linear form of heteroscedasticity using Breusch-Pagan test [Breusch and Pagan, 1979]. We test the null hypothesis that measurement error variances are all equal against the alternative hypothesis that they are a linear function of day, day^2 and state dummy variables. In fact, error variances could potentially be functions of other GIS variables as well, despite the goal here is simply to prove the presence of heteroscedasticity. As Table 1.5 column (4) shows, p-value suggests that sources of measurement error do not have constance variance. Instead, error variances are likely to be fitted values of time and location variables. Therefore, we reject the null hypothesis that measurement error is homoscedastic.
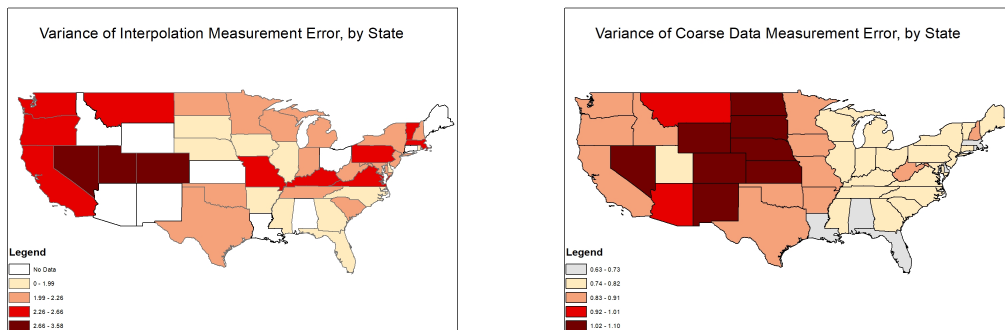
Figure 1.4: σ_u^2 Varies w/ Location

Figure 1.4 shows the state average variance of measurement error for IDW averaging and coarse data mean estimator. In general, states in the west bears larger measurement error noise than states in the east. Variance of measurement error in the southeastern states are the smallest,

most likely because temperature is high and stable throughout the year.
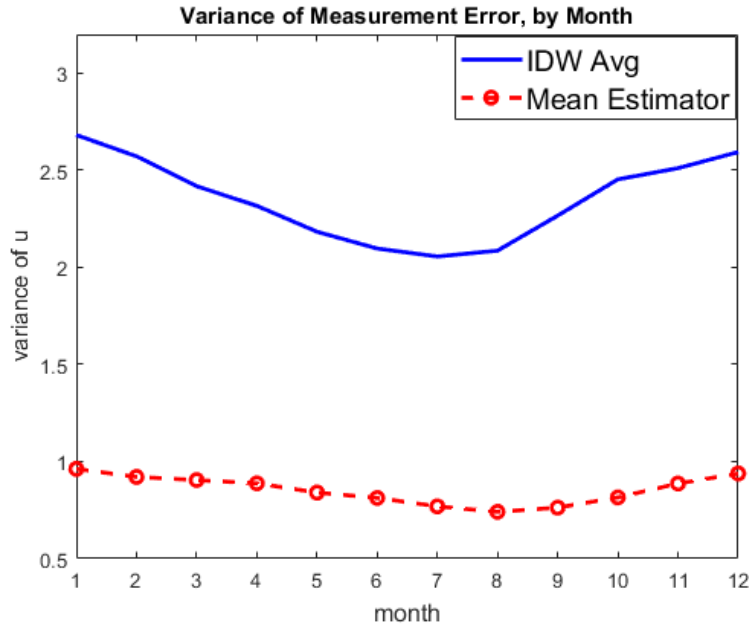
Figure 1.5: σ_u^2 Varies w/ Time

Figure 1.5 shows the monthly average variance of measurement error for IDW averaging and coarse data mean estimator. For both estimators, variances of measurement error demonstrate strong seasonal patterns. Measurement error variances in summer are much smaller than variances in winter. It also confirms that measurement error is highly correlated with true temperature.

3. Normality. To test normality assumption, we use Shapiro-Francia test ([Shapiro and Francia, 1972]). The null hypothesis is that measurement error is drawn from a normal distribution, and alternative hypothesis is that measurement error deviates strongly from a normal distribution. One limitation of Shapiro-Francia test is that it is not valid for large sample size. In other words, the test is always more likely to reject the null with a large sample. Therefore, we sample 5,000 observations each time, and run this normality test for 250 iterations.

Column (5) and (6) from Table 1.5 are the mean test statistics averaged across 250 itera-tions. Large V stats implies large deviation from normal distribution. From the table, it is slightly surprising that PRISM has the largest deviation from normality assumption. Overall, we reject the null hypothesis that measurement error is normally distributed.

4. Independence: Serial Correlation. To test independence assumption requires explicitly defining a dependency structure and testing against that structure. Since measurment errors are generated over time, it is natural to assume serial correlation as an alternative dependency structure. Measurement errors are panel data in our structure so we test serial correlation using Wooldridge's test ([Wooldridge, 2010]).

Wooldridge's procedure in our test involves first-differencing measurement error to remove location-level effect, then regressing measurement error with lagged first-differenced measure-ment error. The test is robust to conditional heteroskedasiticy through clustering regressions at panel level.

In Table 1.5 column (7) and (8), we report estimated coefficient on lagged measurement error and p-value. The p-value implies we should reject the null hypothesis that there is no first order autocorrelation in measurement error. Thus there is indeed serial correlation among measurement error in temperature variables, and independence assumption is rejected.

5. $cov(u_{id}, T_{id}^*) = 0$. To test this assumption, we compute pearson correlation between measurement error and true temperature. Table 1.5 column (9) and (10) report the correlation coefficients and p-value for various sources and types of measurement error. For most types of estimators, our test suggests that measurement error and true temperature are negatively correlated: the higher the true temperature, the smaller the measurement error. The only exception is PRISM,

in which the higher the temperature, the larger the measurement error.

1.4   The Impact of Measurement Error on Parameter Estimates

In section 1.3, with several sets of weather data, we have shown that there are different sources and types of measurement error, which is usually large in magnitude and non-classical in nature. Such measurement error's effect on climate related impact estimation is unclear in several aspects. First, it is unclear how various sources of measurement error interact with each other. For instance, many studies ignore measurement error issue, because they tend to believe aggregation at larger spatial scale could smooth out measurement error introduced at any point location. Second, it is not clear if any one of the several sources of measurement error dominates the other sources. Although we have constructed measurement error empirically in section 1.3.2, the raw weather data are drawn from various sources and cover different sets of locations and time periods, so that comparing across those constructed measurement error is informative but not accurate. Finally, it is not clear what is the impact of the measurement error on existing climate studies.

This section consists of two examples. The first example is an application in the U.S. agriculture context, in which we estimate the link between temperature and yields for 9 major crops. The goal is to show how sensitive the parameter estimates in climate impact studies are to different proxies of temperature. The second exercise is a numerical exercise in which we stack one source of measurement error a time in a logical order to study the interaction effect of multiple sources of measurement error on parameter estimates.

1.4.1 Exercise I: Crop Yield Example

What are the effect of measurement error on actual climate studies? Would the presense of measurement error lead to different conclusions on climate impact studies? To answer these questions, we take the U.S crop yield data and study the impact of measurement error on parameter estimates.

Data and Method

The U.S. crop yield data for the years 1960-2010 are reported by the U.S. Department of Agriculture's National Agricultural Statistical Service. The crop yield is observed at county level and is calculated as county-level production divided by acres harvested. We limit the analysis to counties that can match with our weather data as described in section 1.3.2 and our final dataset includes a total of 2,735 counties. The types of crop used in the analysis include barley, beans, corn, cotton, rice, sorghum, soybean, tobacco and wheat, all of which are major field crops in the U.S. spanning various types of climate regions.

Since the crop yields data are observed at county level, we need to also construct temperature variables at county level. As mentioned in the earlier sections, this implies a need to approximate the temperature at each county centroid. In this application, we pick geographical centroid as the county centroid, then use closest station, simple averaging and IDW averaging to predict the temperature at these centroids. Further, another challenge is that temperature is observed at daily level while crop yields are observed only once a year. There exists a large number of agricultural economics literature that document the relationship between temperature and crop yield, in which different summary statistics of temperature are used. To keep the context simple and general, we use the mean temperature in the growing season[6] as our chosen summary

---

[6]We fix the growing season to months March through October for all crops, though the actual planting date may

statistics and estimate the following regression:

$$\text{Yield}_{it} = \beta \bar{T}_{it} + \varepsilon_{it} \tag{1.7}$$

where $\bar{T}_{it}$ is each county centorid's growing season mean temperature calculated in multiple ways. Besides the growing season mean model, we also put the daily temperature data constructed from the earlier step into 10 temperature bins ($\text{TBIN}_{j,it}$ is the number of days during the growing season in temperature bin j) and estimate the bin regression model:

$$\text{Yield}_{it} = \sum_{j=1}^{10} \beta_j \text{TBIN}_{j,it} + \varepsilon_{it} \tag{1.8}$$

Note that measurement error is not explicitly defined in this example. In an ideal experiment, as is the case in the interpolation measurement error, the true temperature is observed at stations and measurement error can be easily quantified using the leave-one-method as we described in section 1.3. In this example, the temperature that the crops are exposed to are not directly observed and the underlying measurement error is loosely defined to be the difference between any two measures of aggregated temperature, rather than the difference between the true and the approximated temperature. The crop yield example can provide insight into how robust are the estimates of climate impact to different proxies of temperature. Further discussions on the implications of this example are on section 1.4.1.

Results

Estimations from equation (1.7) are reported in table 1.6. In column (1) through (3), the temperature proxies used in the regressions are constructed with daily maximum temperature. In other words, the temperature proxy used is the "county-growing season average of daily maximum temperature". In column (4) through (6), the temperature proxies used are

---

vary depending on crops and local climate.

constructed with the average of daily minimum and maximum temperature—common practice in the empirical studies. There are several important implications from these results. First, the estimated relationship between temperature and crop yield largely depends on the choice of temperature proxy being used. These estimates are different both quantitatively and qualitatively. Comparing models (1)-(3) with models (4)-(6), the levels of the effect substantively vary across the two sets of models. Some of the temperature effects are linear transformations of one set of models from the other set, which implies that if one knows the maximum temperature's effect on crop yields, one could back out the mean temperature's effect on crop yield. Examples of such crop include rice, soybean and wheat. In some other crops, this link between the two sets of models is not obvious. For example, we find that increasing daily maximum temperature by 1°F decreases tobacco yield by about 1.31 bu per acre (column (1)), while using daily mean temperature, the decrease in tobacco yield drops to 0.57 bu per acre and with no significance (column(4)).

Table 1.6: Estimated Temperature's Effect on Crop Yields—Annual Mean Model

	Tmax			(Tmin+Tmax)/2			unit (per acre)	N.obs
	(1) Closest Station	(2) Simple Avg	(3) IDW Avg	(4) Closest Station	(5) Simple Avg	(6) IDW Avg		
Barley	-0.03**	-0.01	-0.02*	0.04**	0.04**	0.04**	bu	40,863
Beans	2.91**	1.98*	2.60**	4.30**	2.57**	3.40**	lb	7,029
Corn	-0.66**	-0.71**	-0.70**	-0.44**	-0.48**	-0.48**	bu	102,949
Cotton	0.56**	0.42**	0.47 **	1.18**	1.00**	1.13**	lb	24,571
Rice	-25.65**	-32.42**	-30.57**	-14.59**	-17.45**	-16.76**	lb	4,554
Sorghum	-0.35**	-0.35**	-0.37**	-0.28**	-0.29**	-0.30**	lb	40,383
Soybean	-0.15**	-0.15**	-0.15**	-0.09**	-0.09**	-0.09**	lb	72,674
Tobacco	-1.31**	-1.48**	-1.38**	-0.57	-1.07**	-0.79*	bu	17,461
Wheat	-0.19**	-0.18**	-0.19**	-0.12**	-0.12**	-0.12**	bu	94,086

Second, the estimated relationship between temperature and crop yield may not be robust to different aggregation methods. In column (1) through (3), county centroid temperature is estimated with closest station, simple averaging and IDW averaging respectively. For some types of crops, the estimated relationship is sensitive to how temperature variable is constructed. The difference in impact estimation is 6.8 lb per acre (26% of the point estimates) for rice;

1.9 lb per acre (32% of the point estimates) for beans; and 0.14 lb per acre (33% of the point estimates) for cotton. There are some crops, such as soybean and wheat, whose yields are more robust to various aggregation methods. A plausible explanation is that these crops are grown in the midwestern plains where the daily variation in temperature are largely canceled out at growing season level. In column (4) through (6), the gap of the estimates across varies aggregation methods narrows for some crops, such as rice; but widens for others, such as beans and cotton.

We also estimate equation (1.8) with all nine types of crops. Temperature bin 60-70°F is left out to avoid multicollinearity. The estimated bin coefficients are plotted in figure 1.6 and each point estimate for bin j should be intepreted as "an additional day with daily mean temperature falling into bin j would increase the crop yield by a magnitude of x". The main takeaway message from this exercise is that using different temperature proxies lead to diverging crop yield response especially at the end bins. For example, having an additional day with mean temperature above 90°F decreases the tobacco yield by 8.25 bu per acre using closest station and decreases the tobacco yield by 16.02 bu per acre using simple averaging. Overall, the estimates are noisier at the end bins due to limited number of observations and more misclassification errors as a result of binning.

Implications from the crop yield example

The crop yield example uses two popular models to estimate the relationship between temperature and crop yield and show that different approximation of temperature would lead to substantially different point estimates for parameters of interest. In an ideal experiment, as is the case of interpolation measurement error, the true temperature is observed and measurement error can be easily quantified. In this example, the temperature that the crops are actually exposed to are not directly observed. The underlying measurement error in our strategy is implicitly defined
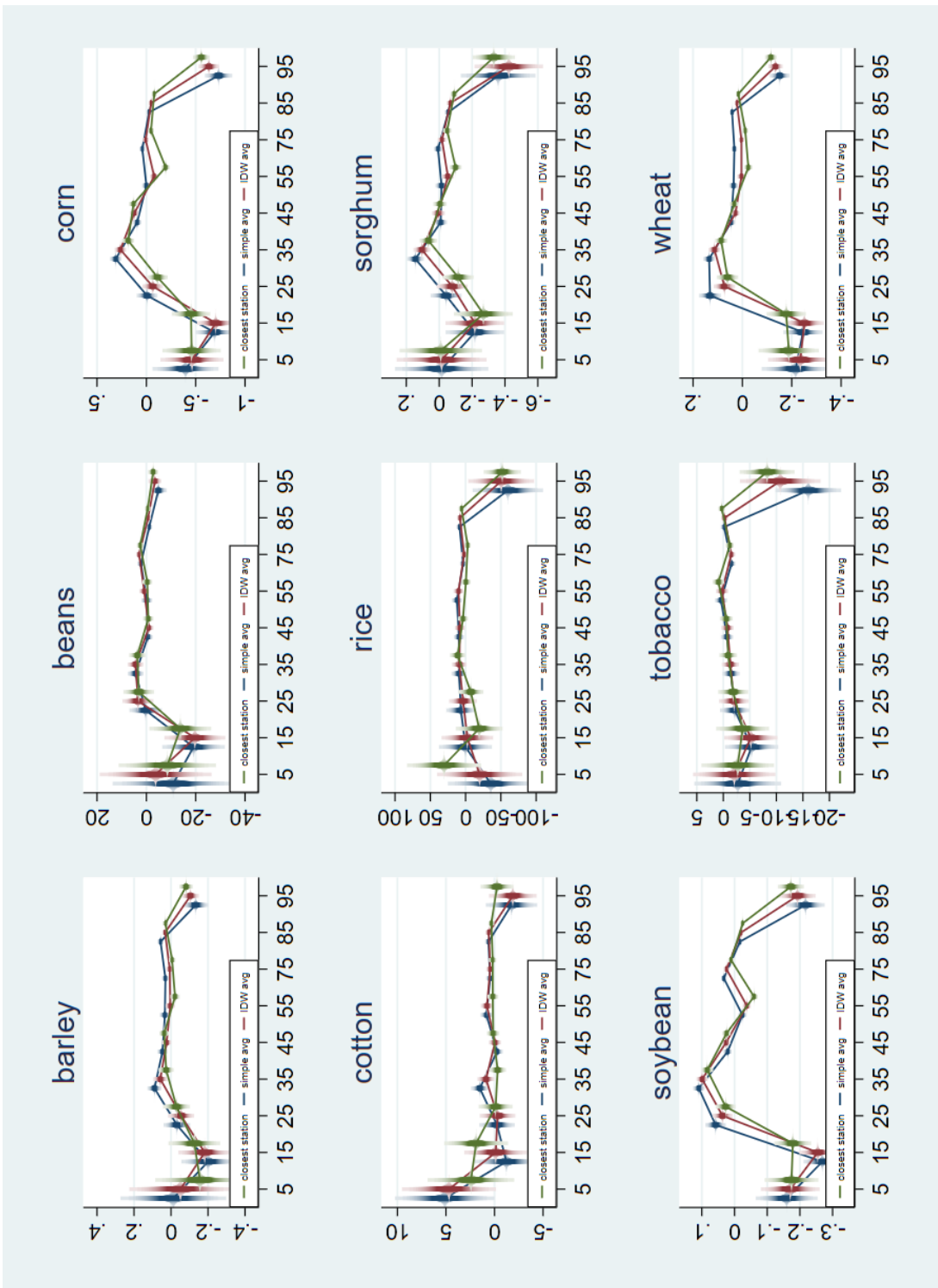
Figure 1.6: Estimated Temperature's Effect on Crop Yields—Bin Model

39

as the difference between any two measures of temperature, rather than the difference between the true and the approximated temperature. The measurement errors implicitly incorporated in this example include three types of measurement errors: Uses of proxy causes divergence between true and assumed DGP is examined when comparing sets of models when temperature is constructed with daily maximum temperature vs. with daily mean temperature; aggregation error is examined by comparing models when temperature is constructed with different interpolation techniques at the county centroid; error due to binning is examined when implementing the bin regression model.

Our crop yield example confirms that the sources of measurement errors are large enough to have significant impact on the identification of temperature response parameters. The impact of measurement errors found in the mean model (equation 1.7) example is conservative. Our strategy is to estimate the impacts of two wrong proxies and compare the difference across the two. These constructed proxies are "wrong" because they bear several sources of measurement error already as we construct them. As a result, these "wrong" proxies are likely closer to each other than to the true value of temperature. Second, the main regressor in the mean model is the growing season mean temperature, in whch variations in daily temperature observations are wiped out. Many climate studies use more complex summary statistics of temperature, such as temperature bins, heating/cooling degree days, which accumulate measurement errors from daily observations. This is another reason why the difference in impact reported in this section is likely conservative. Third, it is standard practice in the literature to add precipitation variable as control in climate impact estimation. The measurement error in precipitation and potentially other weather variables is going to exacerbate the point estimate for temperature's impact. The results from the bin regression model (equation 1.8) is harder to intepete. Our other paper [Carson et al., 2020] show that the bin regression approach produces consistent parameter estimates only under very stringent and highly unlikely assumptions about the true data generating procedure.

1.4.2   Exercise II: Ramona Example

In this subsection, we try to answer these questions: what is the effect on climate impact estimation when there are multiple sources of non-classical measurement error? Would various sources of measurement error cancel out with each other? And are there some sources of measurement error that dominate the others? To answer these questions, we conduct two sets of numerical exercises. In these exercises, we take one monitor as the ground truth, add one source of measurement error a time in a logical order, and observe how climate response function, R-squared, and distribution of measurement error change as we stack various sources of measurement errors together. The numerical exercises are set at Ramona airport weather monitor in San Diego county. Ramona airport weather monitor has high quality hourly temperature measures, and denoted as monitor a. We present two paths for stacking sources of measurement error.

The first path is:

Baseline→closest station→coarse data estimator→aggregation→binning

The second path is:

Baseline→PRISM→aggregation→binning

The second path only adds three sources of measurement error because PRISM introduces both monitor interpolation error and coarse data mean estimation error.

Baseline

Suppose outcome variable is determined by mean temperature at Ramona. We take a simplest possible response function and let the true DGP be on the 45° line and pass through the origin:

$$y = T_{mean}^a$$

Running this regression with ordinary least squares, one should get an R-squared equals to 1 as shown in Figure 3.3, because there is zero measurement error and no idiosyncratic error component.
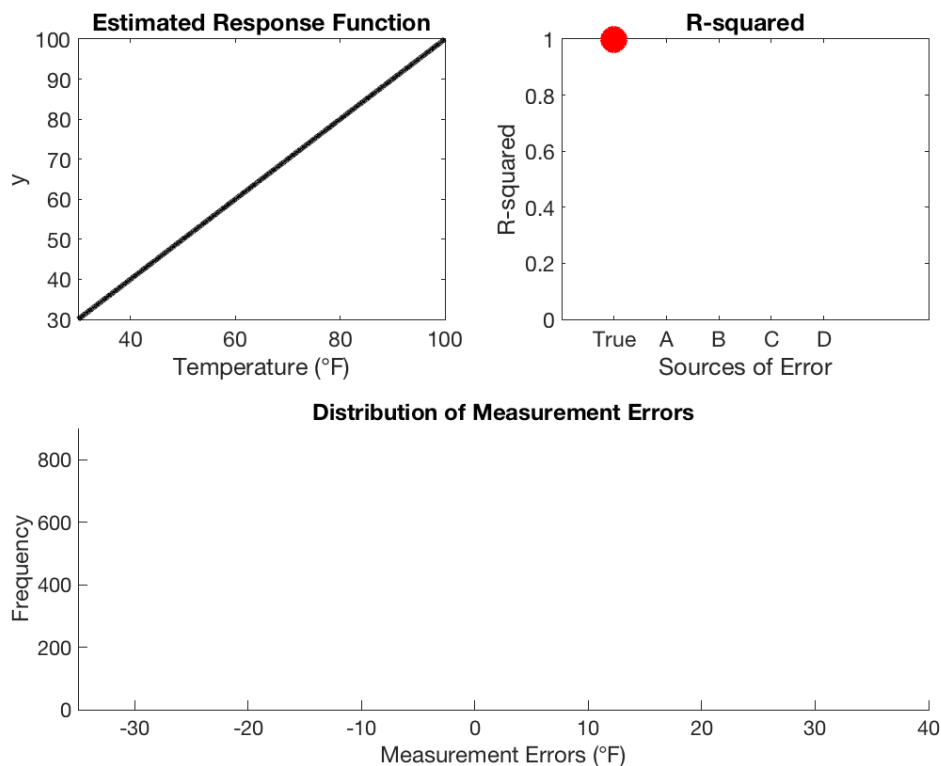
Figure 1.7: Ramona Example: Baseline

Path One

    Type A. Closest Station

Assume that monitor a's reading is not available. We use the weather monitor located at Marine Corps Air Station as the closest station for the interpolation of monitor a. Marine Corps Air Station is about 20 miles away from monitor a, and is denoted as monitor b. The temperature proxy used is T_{mean}^{b}, which is the average of 24 hourly temperature. The model to be estimated is:

$$y = \beta T_{mean}^{b}$$

In Figure 1.8, the slope in the response function is much steeper than the baseline, implying an overestimation in β and the measurement error does not lead to attenuation bias. R-squared drops to 1 from 0.87, which suggests that by interpolating with closest station, 13% of noise to the natural variation in temperature is introduced. The measurement error is asymmetric and centered to the right. Further, the error introduced fails the Shapiro-Francia normality test, so it is distinctly non-normal.
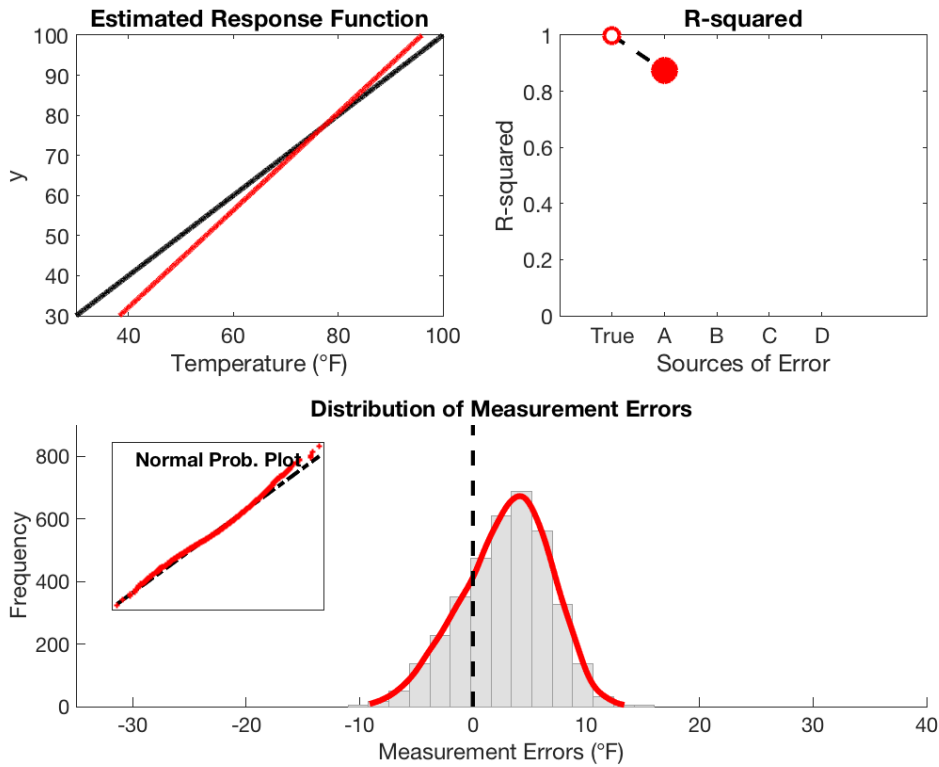
Figure 1.8: Path One: Closest Station

Type B. Coarse Data Estimator

Now suppose instead of reporting hourly temperature, monitor b only reports daily minimum and maximum temperature, which is the case for most NOAA weather monitors. We use the average of daily minimum and maximum temperature as the proxy for daily mean temperature, and estimate this model instead:

$$y = \beta \times (T^b_{min} + T^b_{max})/2$$

In Figure 1.8, R-squred drops from 0.87 to 0.86. By using a statistics that consists of the two extreme order statistics, we introduce another 1% of noise to the natural variation in temperature, which is not large in this instance, but can be a significant source of error elsewhere.

The cumulative measurement error generated by the first two steps gets further away from the center.
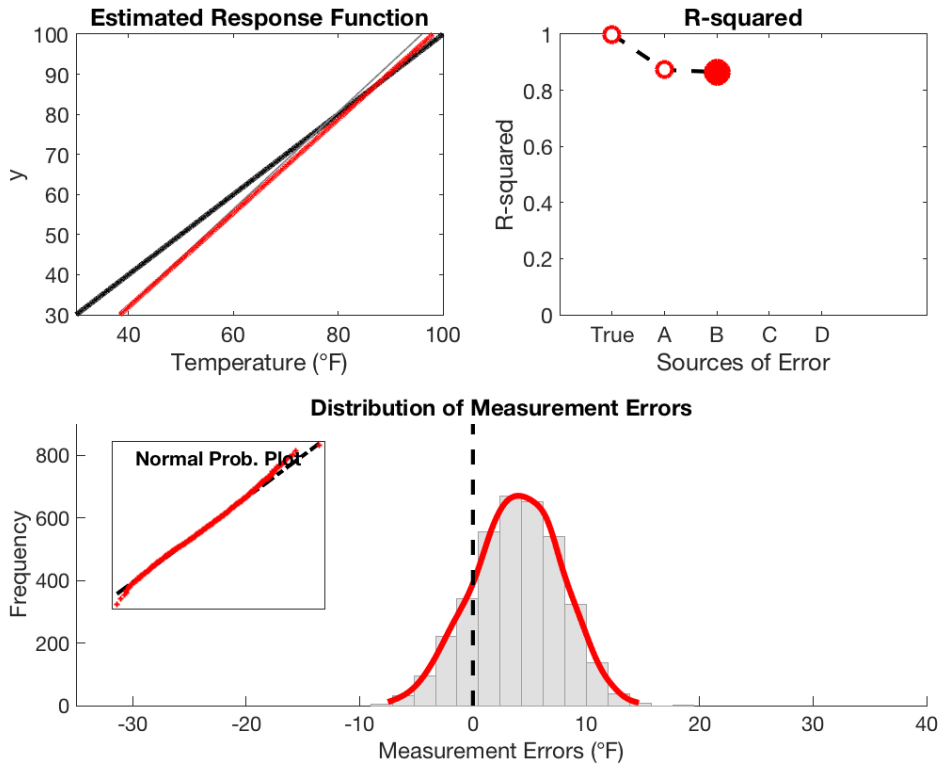
Figure 1.9: Path One: Coarse Data Estimator

Type C. Aggregation

On top of the previous assumptions, suppose in addition, it is not known that the outcome variable is collected from monitor a. The only locational information known about outcome variable is that it is collected from San Diego county. Under this limited information, which is often the case in most empirical studies, the temperature at San Diego county's population centroid is estimated with the closest station method and is used as the temperature proxy for the county. The monitor closest to San Diego's population centroid is located at San Diego Airport,

which is 40 miles away from monitor a, and is denoted as monitor c. The model to be estimated is the following:

$$y = \beta \times (T^c_{min} + T^c_{max})/2$$

In Figure 3.2, the slope in the response function further deviates from the baseline, and the cumulative measurement error at this point clearly leads to an overestimation. R-squared is about 0.83, thus indicating that aggregation may not necessarily alleviate measurement error problems generated by other sources. In other words, sources of measurement error will not magically cancel each other out. The cumulative measurement error generated by the first three steps is substantively left-skewed.

Figure 1.10: Path One: Aggregation

Type D. Binning

Finally, we bin the temperature proxy constructed in the previous step by putting each $(T_{min}^c + T_{max}^c)/2$ into one of the seven bins: below 40°F, 40-50°F, 50-60°F, 60-70°F, 70-80°F, 80-90°F, and above 90°F. For example, for a day with 51°F, $\text{TBIN}_{50\sim60}$ equals 1, and all other bin indicators equal 0. We take 60-70°F as the reference bin, and estimate the following model:

$$y = \sum_{j=1}^{6} \beta_j \text{TBIN}_j$$

Figure 1.11 shows the result of imposing binning specification as the last source of measurement error to temperature. The response function has a kink around 75°F, though in the true DGP response is linear and smooth throughout. R-squared drops from 0.83 to 0.69, suggesting almost another 14% of noise added to the natural variation in temperature due to this binning practice.

Figure 1.11: Path One: Binning

At this final step, the measurement error accounts for about 31% of the natural variation in temperature, which is many times larger than the size of measurement error indicated by [Lobell, 2013] simulation design. The cumulative measurement error is distinctly non-normal, positive, and the response parameter β is estimated with a bias as large as 40% in some intermediate steps, and binning distorts the shape of the response function in this numerical exercise.

Path Two

Path two deviates from path one at the introduction of monitor interpolation error. To interpolate temperature at monitor a, closest station is used in path one, while PRISM reanalysis data is used in path two.

Type A&B. PRISM

PRISM interpolates temperature using information from both weather monitors and GIS variables and reports daily minimum and maximum at high spatial resolution as its primary data product. Therefore, two sources of measurement error — monitor interpolation error and coarse data mean estimator — are almost always automatically introduced whenever one uses PRISM data to estimate mean temperature.

Suppose PRISM data is used to interpolate temperature at monitor a. We find the corresponding grid on PRISM that matches with monitor a's coordinates and take the average of daily minimum $T^a_{min,p}$ and maximum temperature $T^a_{max,p}$ on that grid as the mean temperature proxy for monitor a. The model to be estimated is the following:

$$y = \beta \times (T^a_{min,p} + T^a_{max,p})/2$$

In Figure 1.12, R-squared drops from 1 to 0.9, which suggests that PRISM contributes to about 10% of noise to the natural variation in temperature. Compared to path one, in which R-squared drops to 0.87 after the first two sources are introduced, PRISM does seem to introduce slightly smaller measurement error up to this point. However, when comparing to the estimated response function, PRISM leads to a sizable attenuation bias as the slope for the response function deviates substantively from the 45 degree line. The distribution of measurement error is not centered at zero and fails normality tests.
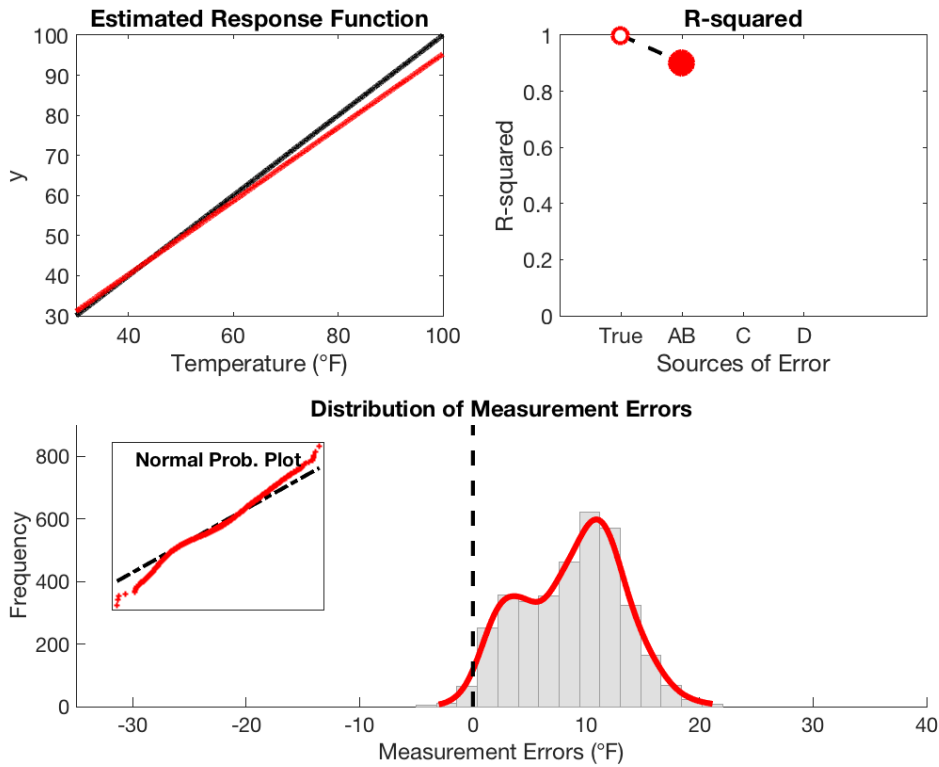
Figure 1.12: Path Two: PRISM

Type C. Aggregation

Like in path one, suppose in addition, it is not known that the outcome variable is collected from monitor a. The only locational information known about outcome variable is that it is collected from San Diego county. To be consistent with path one, we estimate temperature at monitor c with PRISM data. The model to be estimated is the following:

$$y = \beta \times (T^c_{min,p} + T^c_{max,p})/2$$

In Figure 1.13, the slope in the response function is much steeper than the baseline, similar to the slope in path one. This confirms that estimation involving aggregation can lead to bias in unknown directions, depending on the particular aggregation method being applied. The size

and sign of this bias vary with the choices of centroids and the choices of interpolation methods. R-squared drops from 0.90 to 0.84. The cumulative measurement error generated by existing steps is left-skewed.
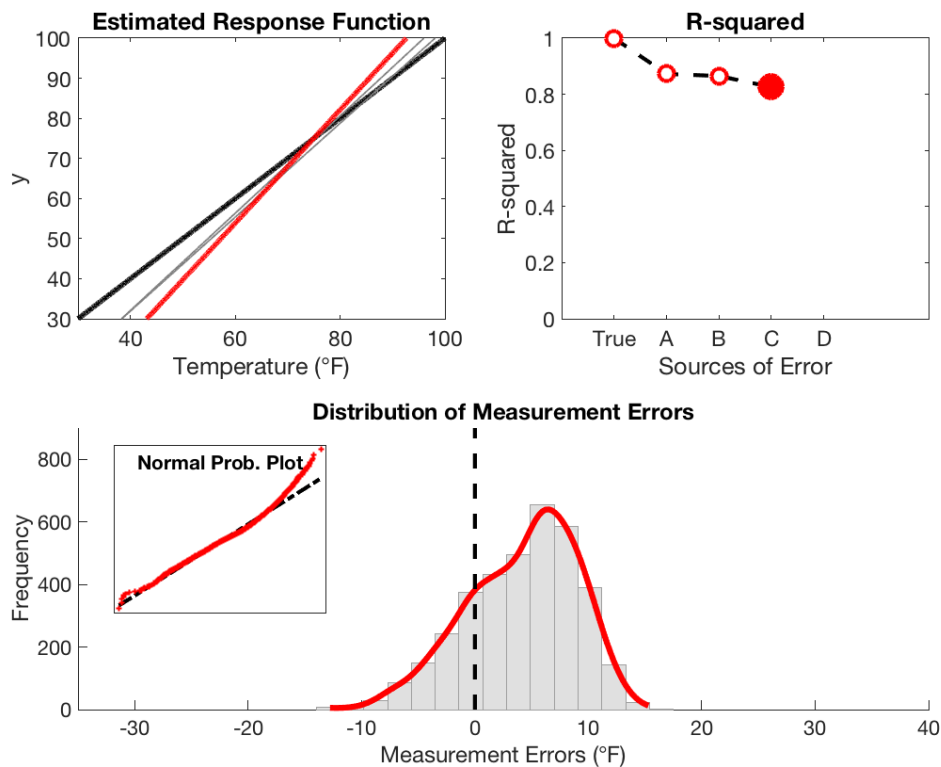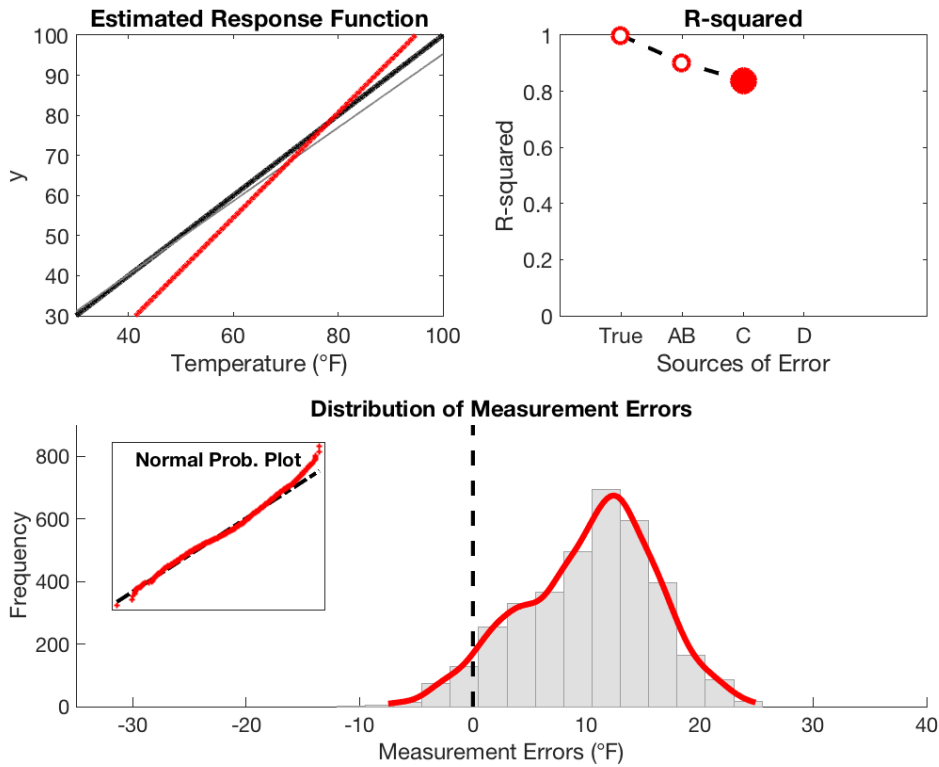

Figure 1.13: Path Two: Aggregation

Type D. Binning

Finally, we bin the temperature proxy constructed in the previous steps, by putting each $(T_{min,p}^c + T_{max,p}^c)/2$ into seven different bins as defined in path one. We still take 60-70°F as the reference bin, and estimate the following model:

$$y = \sum_{j=1}^{6} \beta_j \text{TBIN}_j$$

Figure 1.14 presents the result of imposing binning specification as the last source of measurement error to temperature. Similar to path one, the plotted response function suggests the presense of a kink around 75°F, which is not supported by the true DGP. The R-squared drops from 0.84 to 0.71, suggesting another 13% of noise added to the natural variation in temperature due to the binning practice.
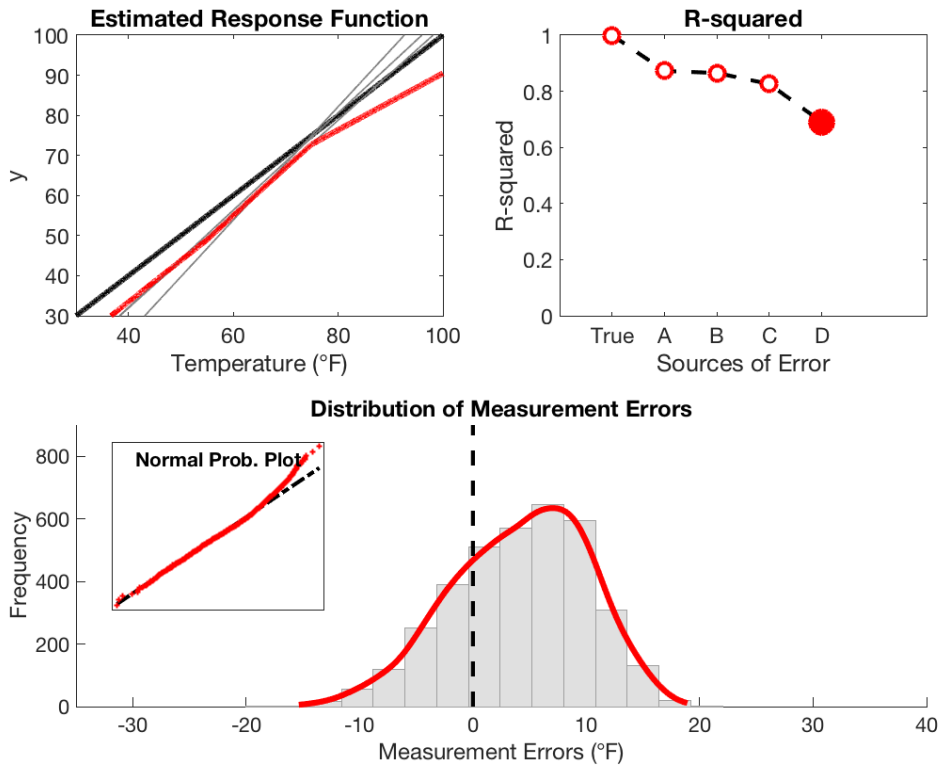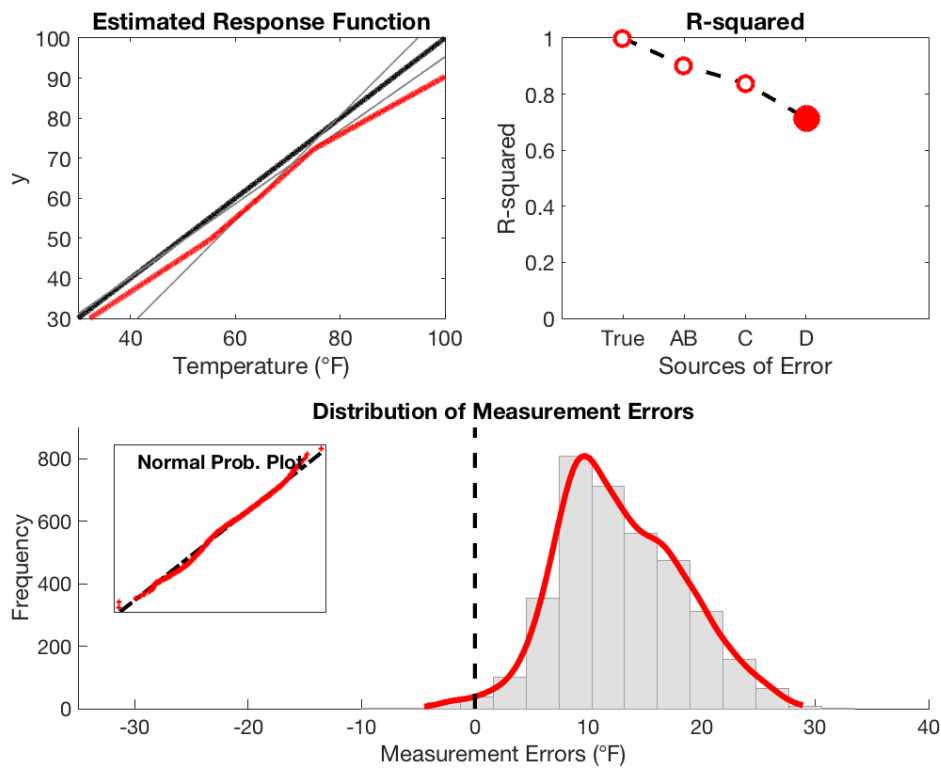

Figure 1.14: Path Two: Binning

At this final step, the measurement error accounts for about 29% of the natural variation in temperature, which is slightly smaller than the size of measurement error generated in path one. The cumulative measurement error is distinctly non-normal, positive, and the response parameter β is estimated with a bias as large as 30% in some intermediate steps, and binning distorts the shape of the response function with a kink at 75°F in this exercise.

Implications from the Ramona Example

Based on the results from these two paths, we find that all four sources of measurement error as characterized in section 1.3 significantly contribute to bias in the estimation of a simple linear climate impact model. The climate impact parameter β is estimated with a bias of as large as 40% in some intermediate steps and in ambiguous directions. Similar to the test results reported in section 1.3.3, the cumulative measurement errors constructed in these exercises are also not normally distributed, not mean zero, thus classical measurement error assumptions would not hold.

Between the two paths, in terms of the R-squared estimated in the final step, using PRISM rather than closest station to interpolate temperature leads to slightly smaller measurement error noise. A plausible explanation is that PRISM data product is spatially smoothed, so there may exist an internal function linking predicted temperature at monitor a and monitor c in the PRISM model algorithm. In other words, the correlation between temperature predicted at monitor a and monitor c can be highly correlated in PRISM's underlying physical model. In contrast, in the closest station method, weight is placed on solely on one monitor for prediction, so that other neighboring monitor's information is not incorporated at all, leading to larger measurement error when one use monitor c to predict temperature at monitor a.

Among the four sources of measurement error, both interpolation and binning contributes to about 10-15% of noise. Coarse data mean estimator contributes to around only 1% of noise to the variation in true temperature, which is relatively small compared to other sources. Aggregation error is harder to intepret: based on the changes in R-squared it contributes to about 4-6% of noise. However, the estimated impact parameter suggests that adding aggregation error changes the point estimates by large. Because aggregation is so oftenly used and in various arbitrary ways,

how the adoption of different aggregation methods affect identification of climate impact models could be one possible extension of this work.

1.5   Correcting for Measurement Error

Existing literature often overlook measurement error issue mainly because they implicitly assume measurement error in weather variables is classical. Under classical measurement error assumption, existing empirical estimations are relatively conservative and subject to, at most, mild attenuation bias. However, as we have tested in section 1.3 and 1.4, we reject the null hypothesis that measurement error is classical using multiple sources of measurement error data. We also show in section 1.4 that the size of measurement error is much larger than people usually expect, which can cumulatively account for 30% of noise in temperature.

In this section, we first show that if one or more classical measurement error assumptions are relaxed, the OLS estimator for impact estimation can be biased in unknown directions, often depending on the correlation between true temperature and measurement error. We then propose a solution to correct for the sources of measurement error as characterized in section 1.3.

There are two common approaches to correct for measurement error bias. The first is an instrumental variable approach. In this approach, people use one measure of temperature as an instrument for another measure of temperature, under the assumption that measurement error generated in these two methods are uncorrelated. However, as we have shown in section 1.3, different sources of measurement error all vary across time and location; also, the smaller the true temperature, the larger the measurement error. In fact, most types of measurement error in temperature variables are not independently distributed. Therefore, we can rule out instrumental variable

54

approach to correct for measurement error in temperature variables. The feasible approach to correct for measurement error is to construct correction factor. This idea is built on the attenuation factor derived under classical measurement error assumptions. As implied by equation (1.6), we can always back out true parameter β as long as there are consistent estimators for σ_{T*}^2 and σ_u^2, if measurement error is classical. Likewise, we can correct for nonclassical measurement error bias using a similar approach.

In section 1.5.1, we derive correction factors for non-classical measurement error model when different classical measurement error assumptions are relaxed. In section 1.5.2, we show the construction of correction factor under different types of outcome variables in a linear framework.

1.5.1   Non-Classical Measurement Error Model

In this section, we characterize cases when some of the classical measurement error assumptions are relaxed. Explicitly writing analytical forms of $\hat{\beta}$ is fundamental to the construction of correction factor.

The model to be estimated is still characterized by equation (1.5). Now let us relax assumptions A1 through A3 as listed in section 1.3. A4 is $E(u_{id}^2) < \infty$, implying that the variance of measurement error is finite, which is not rejected throughout the paper.

Relaxing A1: $u_{id} \sim i.i.d.N(0, \sigma^2), \ \forall i, d$

There are four assumptions embedded in A1, including mean zero, homoscedasiticity, normality and independence. We have shown in section 1.3 that these assumptions are not valid by testing them on various sources of measurement error data. Under some regularity conditions, relaxing each or all of A1 assumptions will always yield the following[7]:

---

[7]Further rationales on the assumptions in A1 can be found in appendix 1.8.4

$$\hat{\beta} \xrightarrow{p} \beta \times \underbrace{\frac{\widehat{var}(T^*)}{\widehat{var}(T^*) + \widehat{var}(u)}}_{\text{Correction Factor}} \tag{1.9}$$

$\hat{\beta}$ converges in probability to the true parameter β multiplied by a correction factor.

Relaxing A2: $cov(T^*_{id}, u_{id}) = 0$

If we relax the assumption that $cov(T^*_{id}, u_{id}) = 0$, $\hat{\beta}$ converges in probability to the following under regularity conditions:

$$\hat{\beta} \xrightarrow{p} \beta \times \underbrace{\frac{\widehat{var}(T^*) + \widehat{cov}(T^*, u)}{\widehat{var}(T^*) + \widehat{var}(u) + 2\widehat{cov}(T^*, u)}}_{\text{Correction Factor}} = \beta \times \underbrace{(1 - \hat{b}_{uT})}_{\text{Correction Factor}} \tag{1.10}$$

where $\hat{b}_{uT}$ is the regression coefficient of estimating u_{it} on T_{it}. Notice that classical measurement error is a special case with $\widehat{cov}(T^*, u) = 0$.

As equation (1.10) shows, the sign and size of $\hat{\beta}$ depend on the regression coefficient $\hat{b}_{uT}$. When $\hat{b}_{uT} \geq 1$, presense of measurement error reverses the sign of the estimated $\hat{\beta}$; when $0 < \hat{b}_{uT} < 1$, $\hat{\beta}$ is biased towards zero; when $\hat{b}_{uT} < 0$, $\hat{\beta}$ amplifies the magnitude of true β. Bias in $\hat{\beta}$ is eliminated when $b_{uT} = 0$, which is a typical assumption of Berkson's error and is not relevant in climate economics context.

Using the same sets of measurement error as constructed in section 1.3, we estimate the empirical bias factors with equation (1.10). In Table 1.4, column (2) presents the empirically estimated attenuation factors when assumption A2 is relaxed. The result shows that interpolation methods and coarse data mean estimator still introduce attenuation bias even when A2 is relaxed,

which is consistent with results from numerical exercise obtained from section 4. Bias gets larger especially for some commonly used interpolation methods, such as closest station and PRISM.

Relaxing A3: $cov(\varepsilon_{id}, u_{id}) = 0$

If we relax assumption $cov(\varepsilon_{id}, u_{id}) = 0$, $\hat{\beta}$ converges in probability to the following:

$$\hat{\beta} \xrightarrow{p} \beta \times \frac{\widehat{var}(T^*)}{\widehat{var}(T^*) + \widehat{var}(u)} + \frac{\widehat{cov}(u, \varepsilon)}{\widehat{var}(T^*) + \widehat{var}(u)}$$

The first term on the right hand side is just the correction factor when assumption A1 is relaxed. The second term captures the correlation between measurement error and idiosyncratic error. An interesting observation is that when $\widehat{cov}(u, \varepsilon) = \beta \widehat{var}(u)$ holds, $\hat{\beta}$ is consistent. This condition suggests that the coefficient of regressing ε_{id} on u_{id} is $\hat{\beta}$ or the effect of measurement error u_{id} on y_{id} is the same as the effect of T_{id}^* on y_{id}. In general, when A3 is relaxed, the measurement error bias could be either positive or negative, and when $\widehat{var}(u) << -\beta \widehat{var}(T^*)$, the sign of $\hat{\beta}$ could also be reversed.

1.5.2 Correction Factor Method: An Illustration

In section 1.5, we have derived correction factors under several non-classical measurement error models. As long as there exist consistent estimates for each component in the correction factor, we can always correct for bias introduced by non-classical measurement error. In this work, we take "relaxing A2: $cov(T_{id}^*, u_{id}) = 0$" from section 1.5.1 as an example to demonstrate how to use correction factor approach to correct for measurement error bias.

As indicated by equation (1.10) under assumption A2, we need consistent estimates of

$\widehat{var}(T^*)$, $\widehat{var}(u)$ and $\widehat{cov}(T^*, u)$ to back out true parameter β. We need to build a prediction model of $\hat{\sigma}^2_{u_{id}}$ for given types of measurement error at any time and location. We have shown how to quantify measurement error at monitors in section 1.3, and these empirical measurement errors can be used as training data to build a prediction model for $\hat{\sigma}^2_{u_{id}}$ at any time and location. Since $T^*_{id} = T_{id} - u_{id}$, T_{id} is known because it is the temperature proxy constructed arbitrarily by researchers, and u_{id} is known by construction, thus T^*_{id} can be derived. Likewise, we can replicate the first step to obtain $\widehat{var}(T^*)$ and $\widehat{cov}(T^*, u)$ and back out the true parameter β with equation (1.10). An equivalent method to estimating those three parameters is to estimate the coefficient $\hat{b}_{uT}$ and then backout β based on equation (1.10) under assumption A2.

In this section, we use an example to illustrate how to use correction factor method to correct for bias in measurement error models. A complex model that is capable of capturing key features of measurement error for any types of geographical locations would involve more data work, which is beyond the scope of the current project. In this small scope exercise, 258 valid weather monitors in Southern California, are randomly partitioned into two parts: 248 monitors are used to fit a Multivariate Adaptive Regression Splines (MARS) prediction model for correction factor; 10 monitors are used as test data to examine the performance of the correction factor approach. We find that this correction factor method works well when measurement error is relatively large. It might "over-correct" for the measurement error bias when measurement error is small in magnitude. A more comprehensive prediction model could potentially improve this correction factor method, and the construction of such prediction model is an important extension of this current project.

Consistent Estimation of Correction Factor

As illustrated in equation (1.10), the key step of the correction factor method is the estimation of the term $\hat{b}_{uT}$, and let us denote it as "CF" for the rest of the subsection to avoid

notation confusion. There are three steps to consistently estimate the term $\hat{b}_{uT}$ for any given location and time, and we describe in detail the steps involved.

1. estimate $\hat{b}_{uT}$ at monitors

2. fit a MARS model to predict $\hat{b}_{uT}$ as a function of GIS and time variables

3. predict $\hat{b}_{uT}$ at given location & time

First, we use closest station method to interpolate temperature at each of the 248 monitors spanning the year 1960-2017. The method is similar to the method described in section 1.3, where we drop one monitor a time and use the remaining monitors to interpolate the temperature at the target monitor. The only difference from section 1.3 is that the 10 monitors used for testing are excluded. After empirically estimate measurement error u_{ita} at each monitor i, at day d for year t, we estimate correction factor CF_{it} seperately for each pair of monitor i and day d with the following regression:

$$u_{idt} = \text{CF}_{id} \times T_{idt} + \varepsilon_{idt}$$

This step yields a total of 73,419 estimates of CF_{id} at monitor-day level, and these estimates are used as the dependent variable in the next step.

Second, we construct a MARS model to predict CF_{id} as a function of the latitude, longitude, and other factors that may influence the magnitude of measurement errors. The MARS model is a flexible non-parametric regression technique, which can be viewed as a piece-wise linear model that automatically estimates nonlinearities and interactions between variables. We fit a group of GIS features that matches these 248 monitors with the MARS model. Table 1.7 is a detailed description of the variables used in the model. The raw GIS data are obtained from DIVA-GIS[8] and the distance to various geographical features are calculated with ArcGIS.

---

[8]http://diva-gis.org/gdata

Table 1.7: Summary Statistics for GIS variables

	Mean	SD	Description
latitude	34.01	(0.81)	
longitude	-117.33	(0.95)	
elevation	620.56	(533.72)	in feet;
distocoast	81.12	(67.33)	in km;
distowater	11.08	(9.23)	in km; distance to permanent water
distolwate	104.18	(55.46)	in km; distance to land water
distow	10.93	(9.22)	in km; distance to any water body
distoproad	12.32	(13.89)	in km; distance to highway
landcove_1			factor variable with 10 levels; landcover type
pop_1	446.21	(996.18)	in km^2; population density
Observations	73419.00		

Figure 1.15 is a variable importance plot based on the fitted MARS model. The model summary and model selection plot are available in section 1.8 Figure and 1.20. The key variables selected in the MARS model are distance to coast, moments of day-of-the-year, elevation, latitude, and landcover types. The generalized R squared of this MARS model is about 0.32. This model can potentially be improved with the inclusion of a larger dataset and richer GIS features.
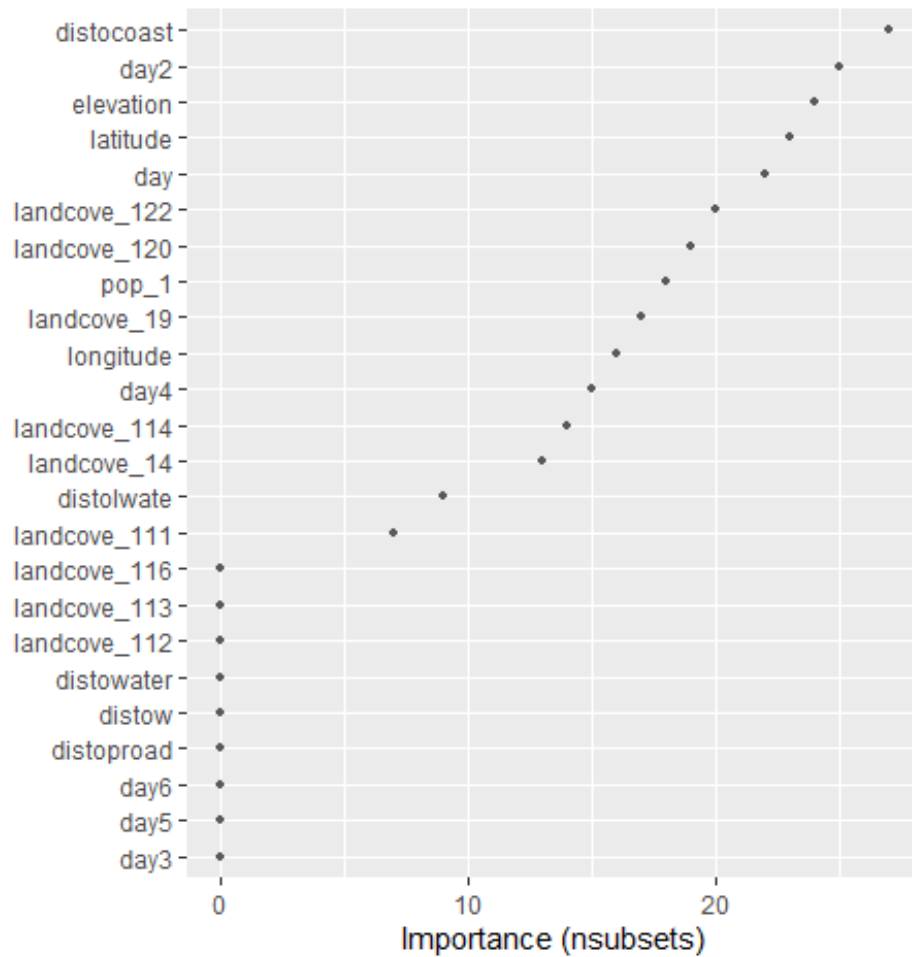
Figure 1.15: Variable Importance Plot from MARS model

Finally, we use the constructed MARS model to predict parameter CF_{id} for each day at each of the 10 testing monitors. The predicted parameters $CF_{id,predicted}$ is the correction factor that will be used to back out the true impact parameter β.

Estimates of impact parameter

To evaluate the performance of the correction factor approach, we estimate a simple linear model at the 10 testing monitors. Let the true data generating process be that $Y_{id} = T_{id}^*$ so that if we estimate outcome variable on true temperature, the impact parameter is 1. We estimate the

following model with the 10 testing monitors:

$$Y_{id} = \hat{\beta}_{OLS} T_{id} + e_{id} \tag{1.11}$$

This model is identified at day-monitor level. T_{id} is the daily maximum temperature on day t at monitor i, predicted with closest station method. Since the true temperatures at these testing monitors are actually known, we can compute measurement error u_{it} at these monitors empirically. We then compare three estimates of impact parameter:

(1) $\hat{\beta}_{OLS}$, which is the coefficient estimated from equation (1.11);

(2) $\hat{\beta}_p = \hat{\beta}_{OLS} \div (1 - \hat{b}^p_{uT})$, in which $\hat{b}^p_{uT}$ is predicted from the fitted MARS model at the testing monitors;

(3) $\hat{\beta}_c = \hat{\beta}_{OLS} \div (1 - \hat{b}^c_{uT})$, in which $\hat{b}^c_{uT}$ is estimated from regressing measurement error u_{id} on predicted temperature T_{id} at the testing monitors.

Figure 1.16 presents impact parameter (1) estimated with no correction; (2) corrected with predicted correction factor; and (3) corrected with true correction factor for each day at two of the 10 testing monitors.
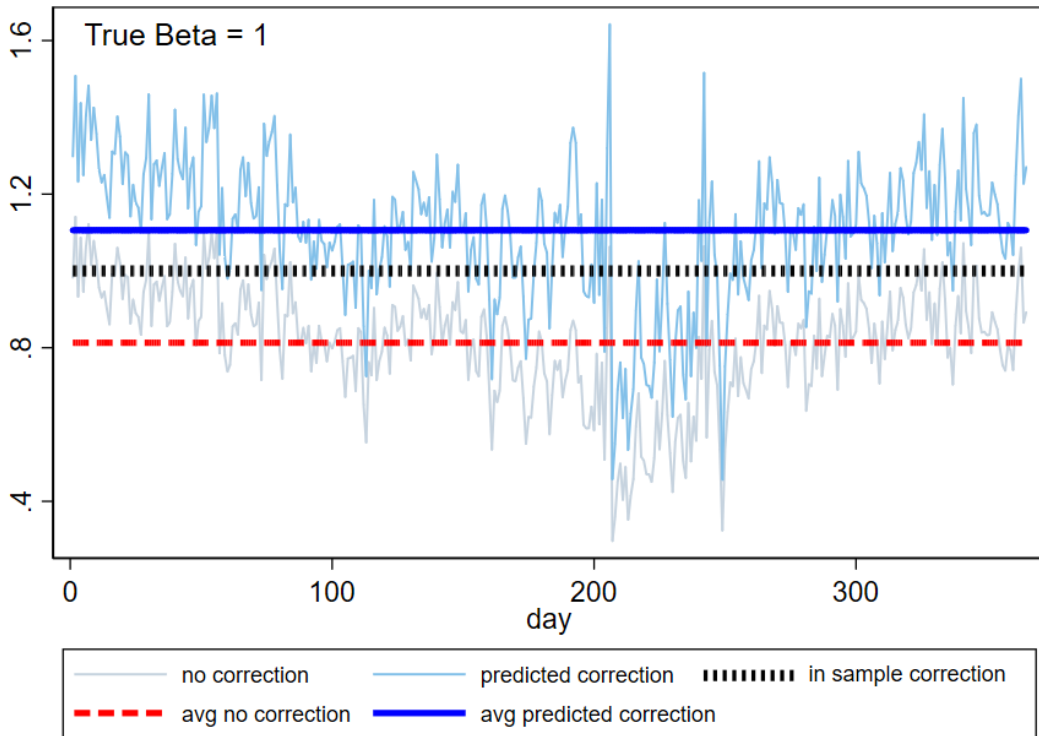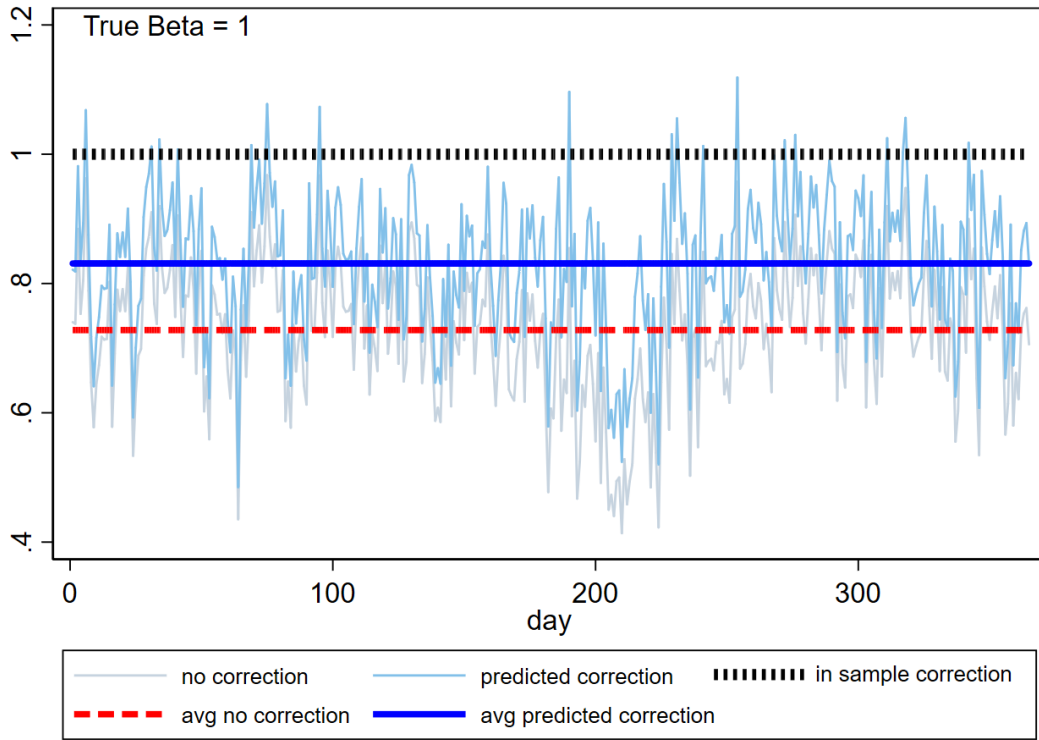
Figure 1.16: Correction for Measurement Error

The key finding is that bias from interpolation measurement error in this area is largely attenuation bias. Our predicted correction factor can "over" or "under-correct" for each specific weather station, but on average the bias is closed by around 30% at the testing monitors. It is likely that a richer prediction model would improve the predicting power and the performance of this correction factor method.

1.5.3   Extension: Using Bootstrap Method to Construct Correction Factor for Aggregation

In section 1.5.2, we have introduced the correction factor method that can consistently estimate correction factor at any given location and time. The location we refered to is a pair of coordinate or a point. However, many outcome variables in economics are actually aggregated to a larger spatial scale, such as city, county, state, or even country. In this section, we illustrate how to apply the correction factor approach to a larger spatial scale.

We use bootstrap technique to estimate the correction factor for a larger spatial scale. Suppose we are interested in estimating the correction factor for a given study region (e.g., city/county/state) at day d. In each iteration k, we sample N points from the study region, get an estimate of the correction factor at each of the N points. To obtain an estimate of the correction factor for day d, we average the correction factors across these N points.

$$\hat{b}_d^k = \frac{1}{N} \sum_{i=1}^{N} b_{id}^k$$

We repeat the above process for K iterations and the county level correction factor is obtained by taking the average across S iterations:

$$\hat{b}_d = \frac{1}{S} \sum_{k=1}^{K} \hat{b}_d^k$$

1.5.4 Discussion on Instrumental Variable Approach

Instrumental variable is a popular approach to correct for measurement error bias. When they are multiple measures of temperature at a given location, we can use one measurement (Z_{id}) as the instrument for the other measurement T_{id}. The first stage regression in IV approach would look like the following:

$$T_{id} = \alpha + \gamma Z_{id} + \nu_{id} \tag{1.12}$$

IV typically work well in dealing with classical measurement error, but when measurement error is non-classical, it is no longer an appropriate approach. In general, three properties must hold to make the IV approach work:

1. Z_{id} is independent of the idiosyncratic error ε_{id} in the response function.

2. Z_{id} is independent of the measurement error in T_{id}.

3. Z_{id} is related to T_{id}

The second condition is unlikely to hold in the climate economics context as we have shown in section 1.3.2 with empirically constructed temperature statistics that measurement error in temperature variables are correlated with the true temperature and that measurement error generally follow seasonal and locational patterns. To be specific, the IV estimator converges to true parameter β multiplied by a ratio. When measurement error in T_{id} is correlated with the instrument Z_{id}, IV is no longer consistent. Further, the size and sign of the measurement error bias depends on the relative size of $\widehat{cov}(u, Z)$.

$$\beta_{IV} \xrightarrow{p} \frac{\widehat{cov}(Y, Z)}{\widehat{cov}(T, Z)} = \beta \times \frac{\widehat{cov}(T^*, Z)}{\widehat{cov}(T^*, Z) + \widehat{cov}(u, Z)} \tag{1.13}$$

It is theoretically possible to construct a Z_{id} that satisfies all three conditions, but it might be hard to implement in practice. One could also find consistent estimate for $\widehat{cov}(u, Z)$ and

$\widehat{cov}(T^*, Z)$, and recover the true β. This would be highly similar to the correction factor approach we have proposed in this section, as both require consistent estimate(s) of the second moments of the measurement error. The correction factor with no IV is more straightforward to implement and more useful from a practical perspective because the IV approach requires the extra step of finding an IV and estimate the covariance between the instrument and the measurement error.

1.6   Conclusion

Climate impact models study how economic agents (e.g., farms, firms or individuals) respond in various ways to exogenous weather stimulus. There are two related modelling issues with such a stimulus. One is the choice of the stimulus variable (e.g., mean, maximum) that causes the observed response. The other issue is the construction of an imperfect proxy for the stimulus variable of interest, because often times the true stimulus variable is unknown or difficult to measure.

Using temperature, a key variable for estimating the impacts of climate change, we explore these two issues. The first comes into play because there is a mismatch between the spatial or temporal scale on which the outcome variable is observed and because commonly used statistics representing temperature tend to be highly correlated. While the second issue, possible measurement error in the stimulus variable, is often noted, it is usually assumed to be "small" and classical (i.e., i.i.d., normally distributed with zero mean). This belief implies that the key parameter estimates of interest are attenuated toward zero and that this attenuation bias is small enough to be ignored after noting that the parameter estimates of interest are likely to be conservative.

In this paper, we use cross-validation method to construct measurement error in temper-

ature variables using various sets of weather data product and show that common assumptions about measurement error do not hold empirically. In fact, measurement error in temperature variables are often large, distinctly non-normal, and vary systematically in predictable ways across space and time.

A major reason behind the divergence between our empirical results and convention wisdom is that the construction of temperature proxy variables almost always involves a series of multiple steps each of which introduces a different source of measurement error. Our study is the first in the climate economics literature to characterize measurement error in temperature variables into four different sources: monitor interpolation error, summary statistics estimation with coarse data, spatial aggregation error, and finally error introduced from binning temperature. Existing literature typically bear at least three sources of measurement error.

We then use two numerical exercises to study the impact on estimation when there are multiple sources of non-classical measurement error. First, we implement an application in the U.S. agriculture context, and find under a simple linear framework that different methods to estimate temperature would lead to as large as 30% difference in the identification of point estimate for some types of crops. Second, we run a numerical exercise with empirical weather and find that adding four sources of measurement error contribute to approximately 30% noise in temperature. Among these sources of measurement error, two common sources — interpolation error and binning — each contributes to at least 10% of noise in temperature. Furthermore, measurement error introduces up to 40% of bias in unknown direction in the estimation of climate impact models.

Finally, we propose an analytical framework to correct for the measurement error bias under non-classical measurement error model, and showcase our correction method with an

example set in Southern California. In particular, we use a correction factor method to correct for the bias in measurement error, and we show how the key statistics of the measurement error distribution for temperature statistics can be obtained in practice and how to use them to construct correction factors to obtain consistent estimates of the parameters of interest.

One of this paper's extension is a modelling system that can produce an estimate of the relevant statistics of measurement error for temperature for a particular interpolation method at any specified U.S. location and date. Applied researchers can use these estimates to construct correction factors in their estimates of temperature impacts. These estimates of measurement error can also be used as one of the building blocks for correction factors in models that involve sources of induced measurement error other than that associated with not having a monitoring station at the economic agent's location. The methodology developed should be extensible to other weather and pollution variables.

1.7 Acknowledgements

Chapter 1 is coauthored with Richard Carson and is being prepared for publication. The dissertation author is the principle researcher on this chapter.

1.8 Additional Figures and Tables

1.8.1 Maps of Weather Stations
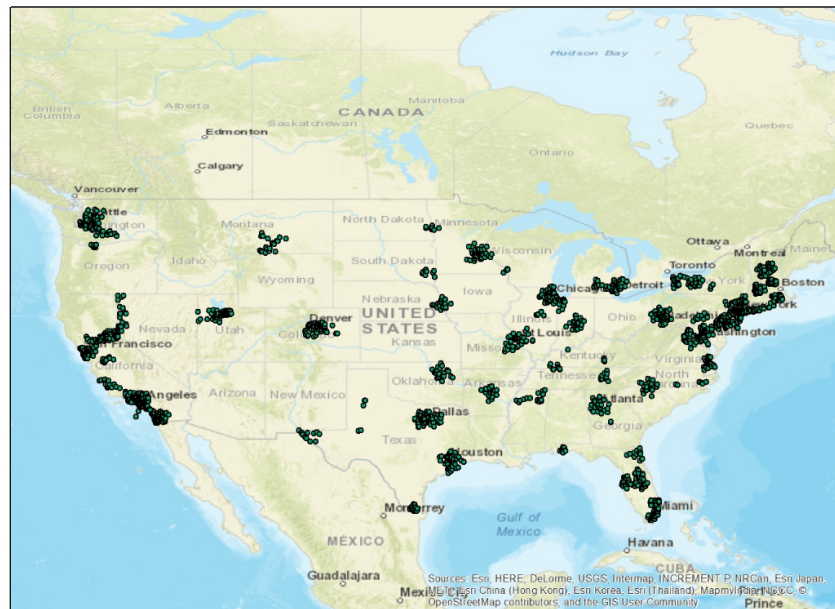
Figure 1.17: Ground weather stations sampled from GHCN

Notes: The map plots the 1516 final stations drawn from NOAA GHCN data set. We use population weighted random sampling procedure to draw 64 MSA, and collect weather data within each of these 64 MSA. This final data set contains stations with at least one other surrounding station within 100km distance.
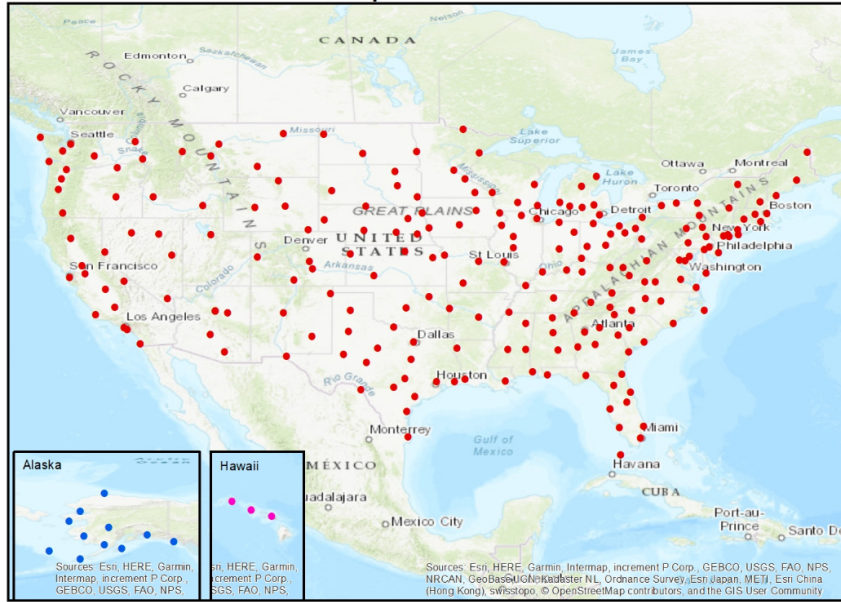
Figure 1.18: Ground weather stations sampled from ISD

Notes: The map plots the 237 airport weather stations drawn from NOAA ISD-LITE data set. PRISM gridded data are also matched to these weather stations except stations in Alaska and Hawaii.

1.8.2 Measurement Error in Mininum Temperature

Table 1.8: Summary Statistics for Monitor Interpolation Errors (Tmin)

	Mean	SD	Min	Max	N.obs
T*	7.19	(10.05)	-47.80	37.80	15,354,109
Closest Station (°C)	2.06	(2.12)	0.00	36.10	15,354,109
Simple Averaging (° C)	1.94	(1.81)	0.00	31.90	15,354,109
IDW Averaging (° C)	1.69	(1.65)	0.00	33.43	15,354,109
Closest Elevation (° C)	2.43	(2.36)	0.00	35.00	15,354,109
PRISM (° C)	1.37	(1.71)	0.00	32.47	10,276,214
N.Stations	31.97	(17.93)	1.00	97.00	15,354,109
Closest Dist (km)	16.76	(10.15)	0.00	94.83	15,354,109
Closest Elev (km)	3.48	(4.41)	1.00	104.00	15,354,109
Elevation (feet)	411.44	(616.84)	0.00	3507.90	15,354,109

70

Table 1.9: Tests on Classical Measurement Error Assumptions (Tmin)

	$\mathbf{E(u)=0}$			Hetero.	Normality		Serial Correlation		$\mathbf{cov(u,T^*)=0}$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Mean(u)	SD(u)	p-value	p-value	V stats	p-value	Lag Corr	p-value	Corr(u,T*)	p-value
Interpolation										
Closest Station	0.02	2.96	0.00	0.00	74.30	0.00	-0.39	0.00	-0.15	0.00
Simple Avg	-0.22	2.64	0.00	0.00	110.77	0.00	-0.37	0.00	-0.25	0.00
IDW Avg	-0.04	2.36	0.00	0.00	135.61	0.00	-0.39	0.00	-0.22	0.00
Closest Elevation	-0.14	3.39	0.00	0.00	93.28	0.00	-0.38	0.00	-0.17	0.00
PRISM	-0.08	2.19	0.00	0.00	227.52	0.00	-0.40	0.00	-0.16	0.00

1.8.3 Numerical Exercise

The result from Ramona example is described in table 1.10

1.8.4 Non-classical Measurement Error Model

Here are some further rationales on each of the four assumptions in A1.

1. $E(u_{id}) = 0$. In equation (1.5), there is a constant term α, which absorbs $E(u_{it})$ under OLS regression, so that e_{it} is always mean zero. $\hat{\alpha}$ is non-zero, and $\hat{\beta}$ converges to true parameter β times the sample analog of attenuation factor, as characterized in equation (1.9).

2. $var(u_{id}) = \sigma^2$. As long as pooled variances of "u_{id}" add up to be the same as σ^2, relaxation of this assumption still yields the same attenuation bias as in classical measurement error case. However, since homoscedasiticity assumption is relaxed, the standard deviation of $\hat{\beta}$ will be different from classical measurement error case.

3. Normality. As long as the alternative distribution has mean zero with variance σ^2, relaxing this assumption also yields the same attenuation bias as in classical measurement error case.

4. Independence. Similar to section 1.3.2, in which we test for independence assumption, we define serial correlation as an alternative dependency structure. The presense of serial

Table 1.10: Ramona Example

	(1) Path One		(2) Path Two	
	$\hat{\beta}$	R-squared	$\hat{\beta}$	R-squared
baseline	1	1	1	1
interpolation	1.21**	0.87	0.92**	0.90
	(0.01)		(0.01)	
coarse data estimator	1.17**	0.86		
	(0.01)			
aggregation	1.41**	0.83	1.31**	0.84
	(0.02)		(0.01)	
binning		0.69		0.71

Impact of an additional day in:	Path One	Path Two	Reference
Bin 40-50	38.62**	49.92**	45
Bin 50-60	48.96**	49.5**	55
Bin 60-70	61.01**	60.62**	65
Bin 70-80	72.81**	72.3**	75
Bin 80-90	79.87**	79.52**	85

correlation in e_{id} has no impact on the attenuation factor as long as pooled variances of "u_{id}" add up to be the same as σ^2 , although it does affect efficiency.

1.8.5 MARS Model

Figure 1.19 is a summary of our fitted MARS model, in which 28 out of 30 predictors are used in constructing this model. R squared is about 0.32. The model has the potential to be improved with a richer date set.

```
Call: earth(formula=c1~., data=newdata, degree=2)

                                     coefficients
(Intercept)                            0.14619528
landcove_111                          -0.32560448
landcove_120                           0.35910497
day4                                   0.00000000
h(28034-distocoast)                    0.00000416
h(distocoast-28034)                   -0.00001201
h(256-elevation)                       0.00004063
h(elevation-256)                       0.00838264
h(36100-day2)                         -0.00001383
h(day2-36100)                         -0.00001237
landcove_120 * pop_1                  -0.00104497
latitude * h(distocoast-28034)         0.00000036
longitude * h(elevation-256)           0.00006968
h(distocoast-28034) * landcove_111     0.00000231
h(distocoast-28034) * landcove_114     0.00000062
h(distocoast-28034) * landcove_122     0.00000278
h(28034-distocoast) * pop_1            0.00000000
h(28034-distocoast) * day              0.00000005
h(28034-distocoast) * elevation       -0.00000003
distocoast * h(day2-36100)             0.00000000
distolwate * h(elevation-256)          0.00000000
landcove_14 * h(256-elevation)         0.00055991
landcove_114 * h(elevation-256)       -0.00014798
landcove_122 * h(elevation-256)       -0.00031390
pop_1 * h(elevation-256)              -0.00000014
h(distocoast-28034) * landcove_19     -0.00000447
h(distocoast-28034) * h(day2-38809)    0.00000000
h(36100-day2) * h(1.22283e+09-day4)    0.00000000

Selected 28 of 30 terms, and 15 of 24 predictors
Termination condition: Reached nk 49
Importance: distocoast, day2, elevation, latitude, day, landcove_122
Number of terms at each degree of interaction: 1 9 18
GCV 0.03346022    RSS 2452.034    GRSq 0.3228082    RSq 0.3240529
```

Figure 1.19: MARS Model Output

Figure 1.20 illustrates the model selection plot that graphs the generalized cross validation (GCV) R-squared based on the number of terms retained in the model. The vertical dashed lined at 28 shows us the optimal number of predictors is 28.

73
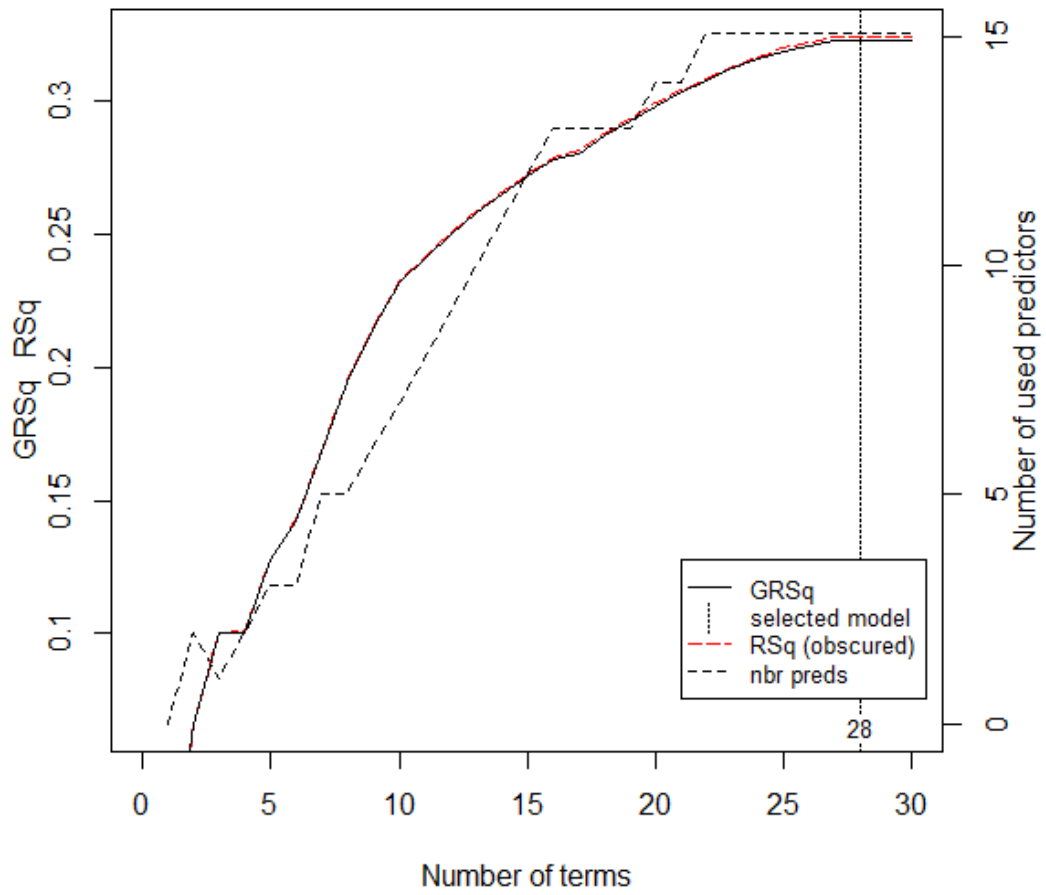
Figure 1.20: MARS Model Selection Plot

Chapter 2

Implications of Using Coarse Data to Approximate Daily Mean Temperature

2.1 Introduction

In recent years, a large number of economics empirical work has been undertaken to study the infuence of climate variables on a wide range of different social and economic outcomes. Regardless of the research design being used, a key methodological step in identifying temperature effect is the measurement of climate variables [Hsiang, 2016]. Daily mean temperature is a key climate variable that has been widely used directly and indirectly in climate economics studies. Daily mean temperature is often used to depict the short-term weather conditions, and to identify heat stress impact on cognitive performance, human conflict etc. Further, it is also used to derive other longer-term temperature-based climate variables, such as heating degree-days, cooling degree-days, count of days in certain temperature ranges etc.

Traditionally, daily mean temperature is computed from two values: the maximum temperature (Tmax) and the minimum temperature (Tmin) for any given 24 hour period. Despite

acquiring temperature data at hourly intervals, most weather stations still report Tmax and Tmin only, thus daily mean temperature today is still computed with Tmax and Tmin. As noted in some recent meteorological studies (e.g., [Bernhardt et al., 2018], [Ma and Guttorp, 2013]), there is a statistically significant difference among different averaging methods, and estimating daily mean temperature with Tmax and Tmin would almost always introduce measurement error.

To explore the potential role of this type of measurement error, I exploit NOAA ground weather station data and PRISM reanalysis data[1]. NOAA-ISD provides hourly temperature data across the U.S, and I select airport weather stations, which have complete hourly observations for 1981-2010. The ground truth temperature is defined as the average of 24 hourly temperatures at these stations. I construct two coarse data mean estimators: (1) estimate daily mean temperature using Tmax and Tmin from ground weather stations, and (2) estimate daily mean temperature using Tmax and Tmin from PRISM — a model-based spatial smoothing data product often used in economics studies.

I quantify measurement error at airport stations by taking the difference between temperatures estimated with those two estimators and the ground truth temperature. The measurement error estimates are then used for three main purposes: (1) to test classical measurement error assumptions, (2) to understand patterns of measurement error across spatial and temporal dimensions, and (3) to estimate bias introduced in climate impact models. The first and second steps provides knowledge on properties of measurement error in temperature variables, which has never been explored before. The third step gauges the importance of measurement error introduced into climate impact models.

I then extend the analysis to measurement error in climate indexes that are derived from

---

[1]PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, created 20 Jun 2019

daily mean temperature. In particular, I explore properties of measurement error introduced in the construction of monthly heating and cooling degree-days, and find that measurement error issue exacerbates in these longer term temperature based climate indexes. Taking weather derivative as an example, I show how measurement error in climate index can potentially affect business risk management decisions through different channels.

This paper extends what is known about the role of measurement error in climate economics. A few climate economics studies mention the issue of potential measurement error in temperature variables in their studies, but since measurement error is assumed "small" and classical, its effect on estimates is considered well understood: small attenuation bias. However, the sources (e.g., interpolation, aggregation, coarse data estimator) and properties (i.e., classical vs. nonclassical) of measurement error and the degree to which measurement error influence estimates of the effects temperature on economics outcomes has not been thoroughly examined. In this paper, I use hourly temperature data acquired at airport weather stations to construct measurment error due to estimate temperature using coarse data (Tmax and Tmin). Moreover, this work also contributes to other fields (e.g., epidemiology, health, recreation and transportation) studying the impact of climate in terms of calling attention to the role of measurement error in temperature variables, and its implications for policy analysis.

The rest of the paper is organized as follows. Section 2.2 provides a literature review on measurement error in climate economics context. Section 2.3 describes procedures to construct measurement error induced by using coarse data to estimate mean temperature, as well as characterizes properties of these measurement error. Section 2.4 characterizes properties of measurement error in other climate index. Section 2.5 concludes this paper.

2.2 Literature Review

Daily mean temperature is a key climate variable that has been widely used directly and indirectly in climate economics studies. It is often used directly to depict short-term weather conditions. For example, [Park, 2018] uses daily mean temperature to study the effect of hot temperature on high-stakes cognitive performance and subsequent educational attainment. [Garg et al., aper] (working paper) document the relationship between temperature and violence in Mexico using daily mean temperature as its key explanatory variable.

Further, daily mean temperature is also used to derive other longer-term temperature-based climate variables. Binned regression categorizes each daily mean temperature observation into pre-determined temperature bins, and counts number of days that fall into each temperature bin. This approach gained prominence in climate economics when used by [Deschênes and Greenstone, 2011] to study the effect of temperature on mortality rate in the U.S. Another popular climate index based on daily mean temperature is degree-days, which accumulates degrees above or below certain threshold temperature over months and seasons. For example, [Burke and Emerick, 2016] use daily mean temperature to construct growing degree-days, and study adaptation to climate change in U.S. agriculture sector. [Considine, 2000] uses heating and cooling degree-days to estimate impact of weather variation on energy and carbon emissions.

Some studies notes the issue of measurement error, but it is typically assumed "small" and classical. This belief implies that the key parameter estimates of interest are attenuated toward zero, and that this attenuation bias is small enough to be ignored. [Lobell, 2013] conducts a simulation exercise on measurement error in temperature and precipitation variables, and finds adding small and classical measurement error in temperature only slightly affects crop yields response estimation. However, the classical measurement error assumption, though often held, is

almost never tested in the literature.

The concept of measurement error is broader than what this paper discusses. In chapter 1, my coauthor and I categorize measurement error in temperature variables into four different sources, following a logical order:

1. Monitor interpolation error

2. Summary statistics estimation with coarse data

3. Spatial aggregation error

4. Error introduced from binning temperature

This work is a natural extension of chapter 1, with a focus on the second source of measurement error. For notation simplicity, I refer to the second source of measurement error as measurement error, though measurement error in temperature have broader meaning. Meteorology studies on summary statistics estimation with coarse data are mostly conducted on mean temperature. A number of papers have compared different averaging methods to derive daily mean temperature. [Bernhardt et al., 2018] find that there is a statistically significant difference among different averaging methods, and estimating certain statistics with coarse data would almost always introduce measurement error. In terms of measurement error in longer run temperature variables, they find that there is significant differences between the two methods of daily temperature averaging on monthly and seasonal time scales. [Ma and Guttorp, 2013] take daily minute average temperature as ground truth, and compare different temperature averaging methods. They find that using hourly temperature as proxy introduces the smallest bias, while using only τ_{min} and τ_{max} leads to the largest bias. In this paper, I take temperature constructed from hourly temperature as the ground truth temperature.

2.3   Characterization of Measurement Error

In this section, I formally characterize properties of measurment error induced by coarse data mean estimator. Section 2.3.1 describes data products used in this paper. Section 2.3.2 describes the procedure for constructing measurement error. Section 2.3.3 provides a simple statistical framework for exploring measurement error in temperature variables. Section 2.3.4 summarizes test procdures and results on classical measurement error assumptions. Section 2.3.5 explores features and stylized facts on constructed measurement error.

2.3.1   Data

This paper uses two data sets on temperature variables: ground weather station data collected from airports and PRISM reanalysis data. Sources are described in more detail in the following subsections.

Ground Weather Station Data

NOAA's Integrated Surface Database (ISD) consists of global hourly and synoptic observations compiled from numerous sources. NOAA ISD-LITE is a simplified version of ISD hourly dataset. I use 237 airport stations across the U.S. for the year 1981-2010. Weather records at airports are relatively accurate because these monitors are better maintained than stations elsewhere. I use NOAA ISD-LITE hourly data to construct measurement error induced by using coarse data to approximate daily mean temperature. Figure 2.1 is a map showing the airport stations we have included in the sample.
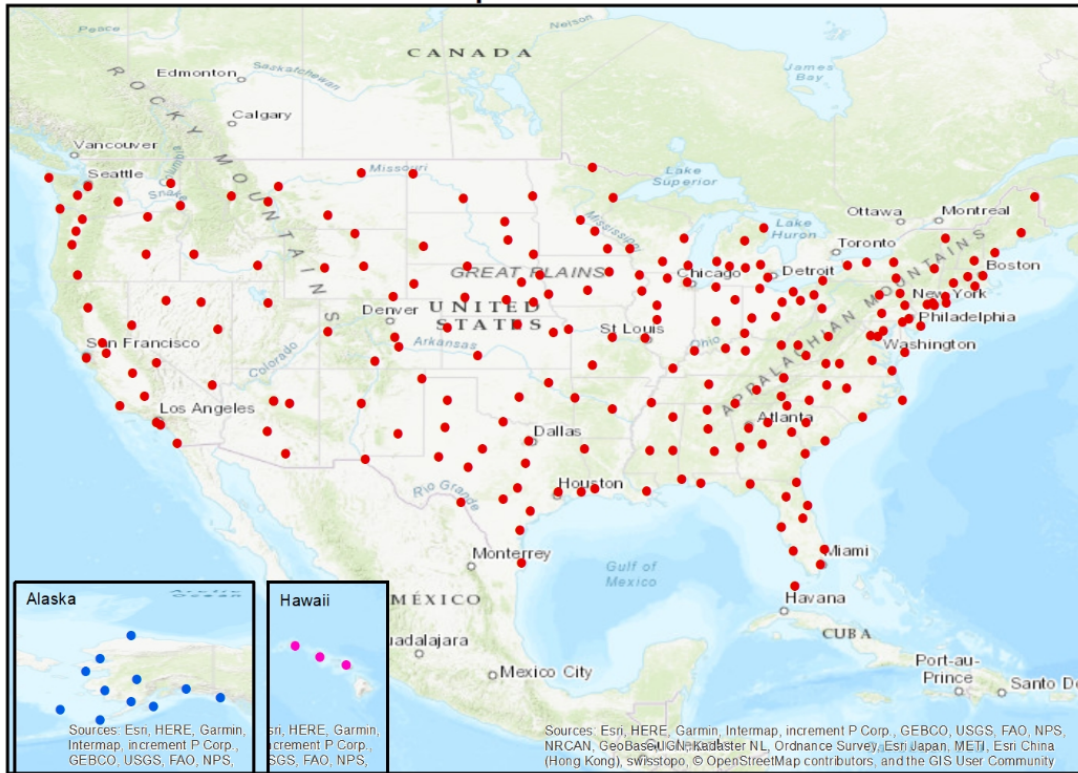
Figure 2.1: Ground Weather Stations Sampled from ISD

Notes: The map plots the 237 airport weather stations drawn from NOAA ISD-LITE data set. PRISM gridded data are also matched to these weather stations except stations in Alaska and Hawaii.

PRISM Data

Over the past decade, there have been large improvements in the measurement of climate variables. Spatial smoothing models bring a combination of climatological and statistical concepts to the prediction of weather variables. An often used weather product is PRISM, which stands for Parameter-elevation Regressions on Independent Slopes Model. It is widely regarded as one of the best geographic interpolation procedures. The PRISM Climate Group gathers climate observations from a wide range of monitoring networks, applies sophisticated quality control measures, and develops spatial climate datasets. It has a fine resolution of 4×4 km grid across the contiguous U.S. PRISM data are widely used in recent climate economics literature, mainly because it is a gridded weather data product that are easy to work with.

2.3.2   Construction of Measurement Error

Estimating summary statistics with coarse data introduces measurement error to temperature variables. A most oftenly used summary statistics of temperature is the daily mean temperature. As noted in some recent meteorological studies (e.g., [Bernhardt et al., 2018], [Ma and Guttorp, 2013]), there is a statistically significant difference between different averaging methods. Daily mean temperature should be computed as:

$$T_{mean} = \frac{1}{S} \int_0^S f(s) ds$$

where S is the number of seconds in a day. The true mean should be an integral over the temperature in each second. However, because many weather stations and PRISM only report Tmax and Tmin, daily mean temperature has to be computed empirically with a much coarser dataset:

$$\hat{T}_{mean} = \frac{\tau_{min} + \tau_{max}}{2}$$

Therefore, measurement error u induced by using coarse data set to estimate daily mean temperature is given by:

$$u = \hat{T}_{mean} - T_{mean}$$

In fact, $\hat{T}_{mean}$ is an unbiased estimator for mean only if τ_{min} and τ_{max} are symmetrically distributed. Apparently, this condition is strange and highly restrictive. In broader statistical sense, to estimate any mean with two extreme order statistics is almost never a good approach.

[Ma and Guttorp, 2013] takes daily minute average temperature as ground truth, and compare different temperature averaging methods. They find that using hourly temperature as proxy introduces the smallest bias, while using only Tmax and Tmin leads to the largest bias. Their result suggests that using hourly temperature should be adequate for representing the true daily temperature. Therefore, I take 24 hours daily mean temperature as the ground truth temperature:

$$\tilde{T}_{mean} = \frac{1}{24} \sum_{h=1}^{24} \tau_h$$

Measurement error in this paper is constructed as:

$$\hat{u} = \hat{T}_{mean} - \tilde{T}_{mean}$$

Table 2.1 is a summary statistics for constructed measurement error using both ground weather station data and PRISM data. "Mean Temp" is the 24 hours daily mean temperature, which is defined in this paper as the ground truth temperature. "Tmin" and "Tmax" are the lowest and highest temperature within 24 hours. $|u|$ denotes absolute value of constructed measurement error, and elevation is measured in meters.

The mean and standard deviation of Tmax and Tmin are close between weather station data and PRISM data. In terms of the magnitude of measurement error, PRISM bears larger measurement error than using weather station. One explanation is that PRISM introduces not only measurement error due to the use of coarse data set, but also interpolation error mechanically introduced in its prediction model.

Size of measurement error caused by coarse data mean estimators varies across space and time. For example, the maximum measurement error generated with weather station occurs in September 1981 at Winslow, AZ. The 24 hours daily mean temperature is about 13°C. However,

Table 2.1: Summary Statistics for Measurement Error

	Mean	SD	Min	Max
Mean Temp (°C)	12.50	(10.89)	-47.13	41.23
Elevation (m)	378.43	(472.02)	0.30	2296.10
Weather Station:				
Tmin (°C)	7.56	(10.71)	-50.60	36.10
Tmax (°C)	18.52	(11.56)	-43.80	50.00
$|u|$(°C)	0.77	(0.67)	0.00	22.86
PRISM:				
Tmin (°C)	7.62	(10.82)	-44.77	30.20
Tmax (°C)	18.77	(11.35)	-31.56	43.78
$|u|$(°C)	1.47	(1.35)	0.00	38.67
Observations	2,280,749			

because lowest temperature drops to -45°C at some point, using coarse data mean estimator leads to a daily mean temperature of -10°C, thus leads to measurement error as large as 22.86°C as reported in the Table 2.1.

2.3.3   Measurement Error Framework

Suppose we wish to estimate coefficient β in the response function:

$$Y_{it} = \beta T_{it}^* + \varepsilon_{it}$$

Following a standard additive measurement error framework, we only observe:

$$T_{it} = T_{it}^* + u_{it}$$

where u_{it} is the measurement error for location i at time t. Classical measurement error assumption usually implies the following conditions:

$$u_{it} \sim i.i.d \; N(0, \sigma^2), \; \sigma^2 < \infty$$

$$cov(u_{it}, T_{it}^*) = 0 \text{ and } cov(u_{it}, \varepsilon_{it}) = 0$$

Under these conditions, $\hat{\beta}_{CME}$ converges in probability to β multiplied by an attenuation factor:

$$\hat{\beta}_{CME} \xrightarrow{p} \beta \times \underbrace{\frac{\sigma_{T^*}^2}{\sigma_{T^*}^2 + \sigma_u^2}}_{\text{Attenuation Factor}} \tag{2.1}$$

As equation (2.1) implies, under classical measurement error assumption, presence of measurement error merely attenuates the estimated impact coefficient toward zero. This attenuation bias has not been a major concern in the literature, as it merely suggests the climate impact identified from these models is conservative.

We show in section 2.3.4 that measurement error in temperature variable is correlated with true temperature, thus measurement error in this paper's context is nonclassical in nature.

Assume in addition that the measurement error is nondifferential (e.g., the measurement error is uncorrelated with ε_{it}), $\hat{\beta}_{NCME}$ converges in probability to:

$$\hat{\beta}_{NCME} \xrightarrow{p} \beta \times \frac{\sigma_{T^*}^2 + \sigma_{T^*,u}}{\sigma_{T^*}^2 + \sigma_u^2 + 2\sigma_{T^*,u}} \tag{2.2}$$

The size and sign of $\hat{\beta}_{NCME}$ depends on the correlation between measurement error and true temperature. The bias factor term can be empirically estimated with $1 - \hat{b}_{u,T}$, where $\hat{b}_{u,T}$ is the regression coefficient from regressing measurement error u on observed temperature T. Using constructed measurement error, I calculate bias factor by estimating $1 - \hat{b}_{u,T}$ under nonclassical measurement error assumptions in later sections of the paper. These bias factors are useful to understand the degree to which measurement error affects OLS results in a simple one regressor

setting.

2.3.4   Test on Classical Measurement Error Assumptions

In this section, I describe test procedures and results on classical measurement error assumptions, including mean zero, normality, homoscedasiticy, independence, and independence from latent variable assumptions. The only assumption I cannot test with weather data alone is the assumption that $cov(u_{it}, \varepsilon_{it}) = 0$, which involves knowledge of idiosyncratic error term ε_{it}. In fact, one can easily test this assumption using a Pearson correlation test with any outcome variable data.

I test five classical measurement error assumptions with both ground weather station and PRISM data. The test statistics is reported on Table 1.5.

Table 2.2: Tests on Classical Measurement Error Assumptions

	$E(\mathbf{u}) = \mathbf{0}$			Hetero.	Normality		Serial Correlation		$cov(\mathbf{u}, \mathbf{T}^*) = \mathbf{0}$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Mean(u)	SD(u)	p-value	p-value	V stats	p-value	Lag Corr	p-value	Corr(u,T*)	p-value
Weather Station	0.54	0.86	0.00	0.00	79.94	0.00	-0.46	0.00	-0.08	0.00
PRISM	-0.20	1.98	0.00	0.00	61.47	0.00	-0.43	0.00	0.00	0.00

1. $E(u_{it}) = 0$. I test mean zero assumption with a simple t-test on the sample mean of measurement error. In Table 2.2, column (1) through (3) are relevant test statistics for the t-test. The p-value indicates that we can safely reject the null hypothesis that measurement error is mean zero at a significant level of 5%.

As equation (2.1) and (2.2) imply, it is the variance and covariance terms of measurement error, rather than the level of measurement error, that determines the size of bias under classical and nonclassical measurement error assumption. Therefore, Var(u) and Cov(u,T*) are more informative of the size of bias caused by measurement error than the mean. Comparing the

86

standard deviation of measurement errors, noise in weather station data is twice as small as noise in PRISM data, which generally suggests that mean estimator using PRISM data set is likely to have larger impact than mean estimator using weather station data set. This finding is well explained by the fact that PRISM introduces a mixture of two measurement errors: coarse data mean estimation error and interpolation error.

2. $var(u_{it}) = \sigma^2$. It is often assumed that measurement error is homoscedastic, but a heterscedasticity test shows this is not the case. In fact, variance of measurement error can be correlated with time, space, and other GIS variables. To test homoscedasticity assumption, I detect a linear form of heteroscedasticity using Breusch-Pagan test. I test the null hypothesis that measurement error variances are all equal against the alternative hypothesis that they are a linear function of day, day^2 and state dummy variables. In fact, error variances could potentially be functions of other GIS variables as well, despite the goal here is simply to prove the presence of heteroscedasticity. As Table 1.5 column (4) shows, p-value suggests that measurement error does not have constance variance. Instead, error variances are likely to be fitted values of time and location variables. Therefore, I can reject the null hypothesis that measurement error is homoscedastic.

3. Normality. To test normality assumption, I use Shapiro-Francia test. The null hypothesis is that measurement error is drawn from a normal distribution, and alternative hypothesis is that measurement error deviates strongly from a normal distribution. One limitation of Shapiro-Francia test is that it is not valid for large sample size. In other words, the test is always more likely to reject the null with a large sample. Therefore, I sample 5,000 observations each time, and run this normality test for 250 iterations.

Column (5) and (6) from Table 2.2 are the mean test statistics averaged across 250 iterations. Large V stats implies large deviation from normal distribution. Overall, we reject the null

87

hypothesis that measurement error is normally distributed.

4. Independence: Serial Correlation. To test independence assumption requires explicitly defining a dependency structure and testing against that structure. Since measurement errors are generated over time, it is natural to assume serial correlation as an alternative dependency structure. Measurement errors are panel data in our structure so I test serial correlation using Wooldridge's test.

Wooldridge's procedure involves first-differencing measurement error to remove location-level effect, then regressing measurement error with lagged first-differenced measurement error. The test is robust to conditional heteroskedasiticy through clustering regressions at panel level.

In Table 2.2 column (7) and (8), I report estimated coefficient on lagged measurement error and p-value. The p-value implies rejecting the null hypothesis that there is no first order autocorrelation in measurement error. Thus there is indeed serial correlation among measurement error in temperature variables, and independence assumption is rejected.

5. $cov(u_{it}, T_{it}^*) = 0$. To test this assumption, I compute Pearson correlation between measurement error and true temperature. Table 1.5 column (9) and (10) report the correlation coefficients and p-value. For ground weather data set, measurement error and true temperature are negatively correlated: the higher the true temperature, the smaller the measurement error. In PRISM, the correlation between true temperature and measurement error is 0.0032, which is extremely small and positive.

2.3.5 Stylized Facts on Measurement Error

The homoscedasiticity test conducted in section 2.3.4 shows that measurement error varies across location and time. In this section, I further explore features of measurement error, especially on spatial and seasonal dimensions. I report results using weather station data only, because PRISM introduces various types of measurement error all at once, which is beyond the scope of this work.

1. Coarse data mean estimator tends to overestimate daily mean temperature.

To estimate relative weights of Tmax and Tmin on true daily mean temperature, I run the following regression[2] for stations in each state seperately:

$$\tilde{T}_{mean} = \alpha_1 \tau_{min} + \alpha_2 \tau_{max}$$

where $\tilde{T}_{mean}$ is the ground truth daily mean temperature defined in section 3.1, and τ_{min}, τ_{max} are Tmax and Tmin for each day. Figure 2.2 is a plot of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ for each state. Hawaii is the only state where daily mean temperature is more determined by Tmax. All other 49 states place more weight on Tmin. Therefore, using the average of Tmax and Tmin would always overestimate temperature for most part of the U.S.

---

[2]There is no constant in this regression to ensure weights are distributed only between Tmax and Tmin.
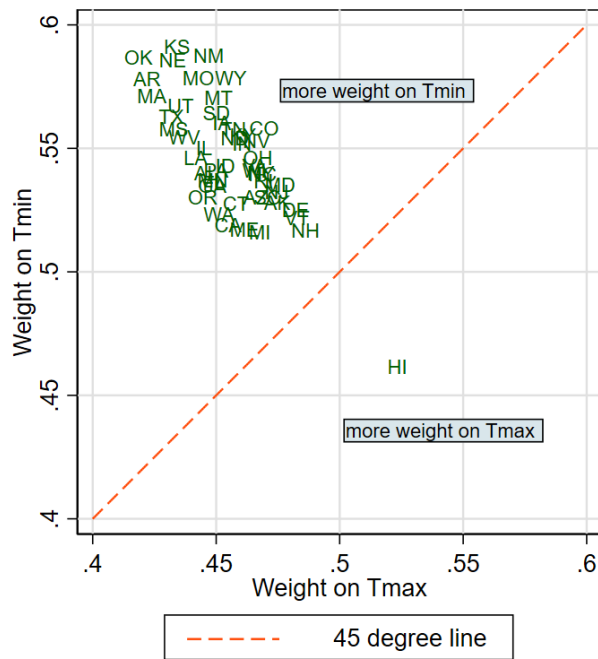
Figure 2.2: Coefficients on Tmax and Tmin, by State

Notes: This plot shows the weight of Tmax and Tmin used for predicting true daily mean temperature for each state. Weights are derived from regressing true daily mean temperature on Tmax and Tmin without constant term. States on the RHS of 45 degree line place more weight on Tmin, and Hawaii is the only state on the RHS which places more weight on Tmax.

2. Measurement error varies across location.

Figure 2.3 (a) and (c) show the state average measurement error using ground weather station data. In terms of the magnitude of measurement error, all states bear positive measurement error. In other words, using Tmax and Tmin to approximate mean tends to overestimate daily mean temperature. In general, states in the western U.S. bear larger measurement error noise than states in the east.
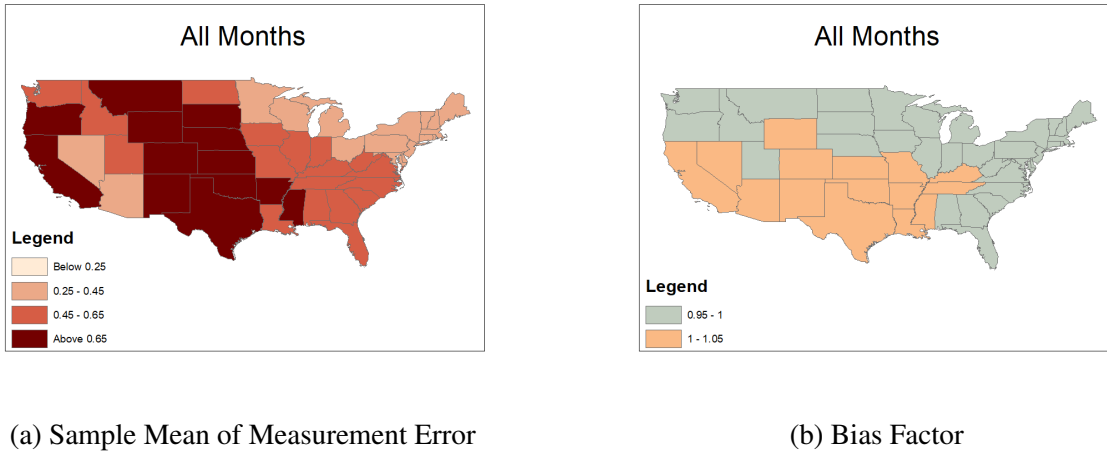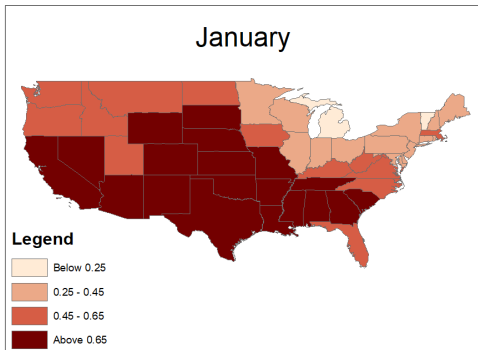
(a) Sample Mean of Measurement Error   (b) Bias Factor

Figure 2.3: Measurement Error Statistics, by State

Notes: These maps plots the spatial distribution of measurement errors in daily mean temperature, averaged over the period from 1981 to 2010 for stations in each state. Ranges in the legend are in terms of degrees Celsius. Subfigure (a) plots the sample mean of measurement error within each state. Subfigure (b) plots bias factors calculated from $1 - \hat{b}_{u,T}$ which is described in section 2.3. No bias implies bias factor equals to 1.

I calculate the fraction term in equation (2.2) for each state seperately. With no measurement error, this fraction term should equal to 1. Under classical measurement error assumption, this fraction term should shrink towards zero as variance of measurement error goes up. When classical measurement error assumption is relaxed, this fraction term could either be above or below 1, depending on the second moment statistics of measurement error.

Figure 2.3 (b) shows the bias factor for each state. There is a clear pattern on how measurement error affects climate impact estimation for various locations: parameters estimated for states in the north and east tend to suffer from mild attenuation bias; parameters estimated for states in the south and west tend to be overestimated. This figure also provides empirical evidence on the fact that when measurement error is not classical, climate impact can be overestimated, which rejects the conventional wisdom that measurement error in climate variables leads to attenuation bias. Besides, this figure also confirms that the magnitude of measurement error does not influence climate impact estimation.
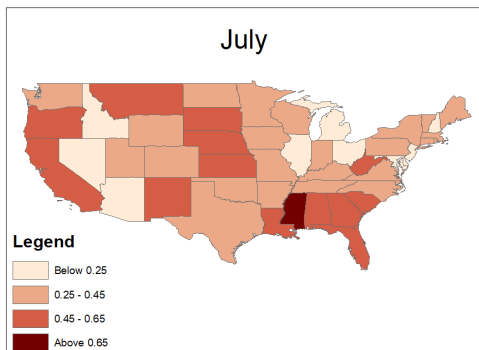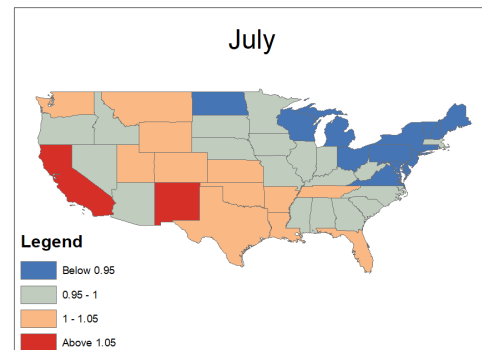
3. Measurement error varies across season.

(a) Jan Sample Mean of Measurement Error
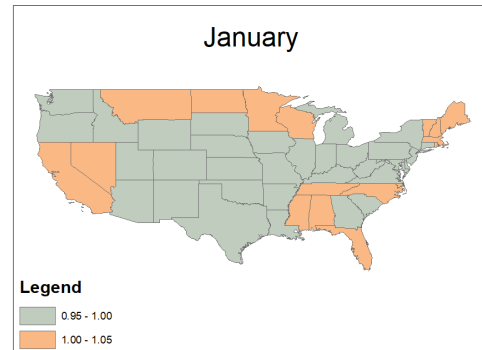
(b) Jan Bias Factor

(c) Jul Sample Mean of Measurement Error

(d) Jul Bias Factor

Figure 2.4: Measurement Error Statistics, by State and Month

Notes: These maps plot the spatial distribution of measurement errors in daily mean temperature for January and July, averaged over the period from 1981 to 2010 for stations in each state. Ranges in the legend are in terms of degrees Celsius. Subfigure (a) (c) plots the sample mean of measurement error within each state in January and July. Subfigure (b) (d) plots bias factors in January and July calculated from $1 - \hat{b}_{u,T}$ which is described in section 2.3. No bias implies bias factor equals to 1.

Figure 2.4 (a) and (c) shows state average mean of measurement error for January and July. Comparing Figure 2.4 (a) and (c), one can see that measurement error in January is much larger than that in July than most locations. A plausible explanation is that hourly temperature is

relatively high in summer, so that Tmax and Tmin are more symmetrically distributed than they are in winter.

There is a clear pattern on how measurement error affects climate impact estimation for various locations: parameters estimated for states in the south and west tend to be overestimated. This figure also provides empirical evidence on the fact that when measurement error is not classical, climate impact can be overestimated, which rejects the conventional wisdom that measurement error in climate variables leads to attenuation bias. Besides, this figure also confirms that the magnitude of measurement error does not influence climate impact estimation.

Figure 2.4 (b) and (d) shows the bias factor for each state in January and July. In January, parameters estimated for the majority of states tend to suffer from mild attenuation bias. In July, parameters estimated for many northeastern states bear an attenuation bias of around 10%. Parameters estimated for states in Southwest have overestimation problem.

It is worth noting that California bear overestimation bias in all seasons. Therefore, studies using California as data source should be careful with result intepretation, especially since data from California is so often used in climate impact estimation studies in various fields (e.g., agriculture, labor productivity, mortality). It is plausible that any large climate impact found can be explained by measurement error in daily mean temperature.

To further explore the seasonality of measurement error, I explore how measurement error varies across month. I regress level and standard error of daily measurement error on months using July as the base level. Standard error is calculated for each station, month and year. Figure 2.5 (a) shows how the magnitude of measurement error varies across months. The solid line in orange shows coefficients on months for all states in the sample. Each point on the line represents

how magnitude of measurement error changes relative to the level in July. Measurement error reaches peaks in March and October, and drops to a trough in early summer. The lines in green and navy represents sets of coefficients for California and Texas respectively. Climate in locations like California has smaller seasonal variation than that in Texas, which is also reflected in their measurement errors' response to seasons: Texas's measurement error varies more with season, and California's measurement error varies less with season.
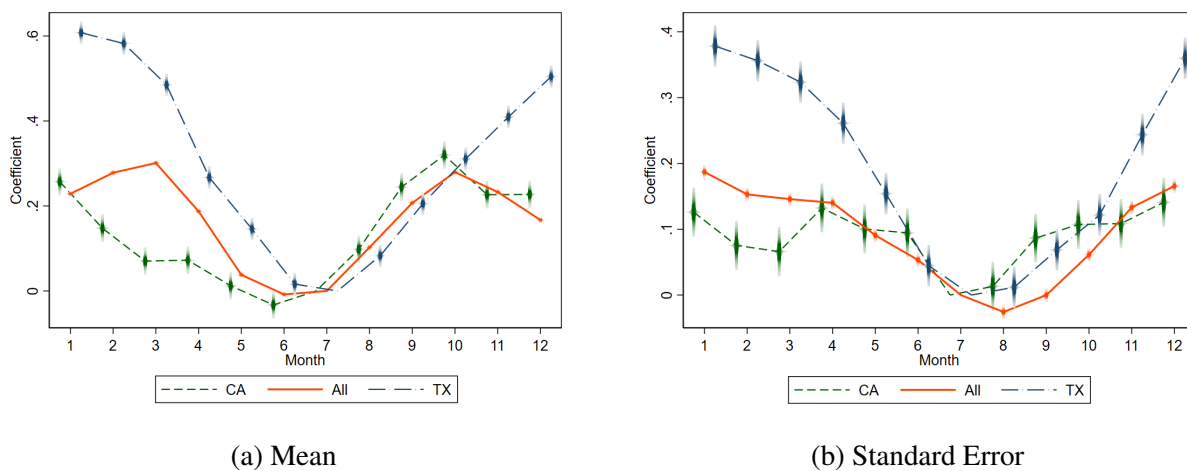

(a) Mean   (b) Standard Error

Figure 2.5: Seasonal Patterns of Measurement Error

Notes: These figures plot the temporal distribution of measurement errors in daily mean temperature. Coefficients in subfigure (a) are derived from regressing measurement error in daily mean temperature on month dummies, with July as the base level. Coefficients in subfigure (a) are derived from regressing standard error of measurement error within each month on month dummies, also with July as the base level. All coefficients are in terms of degrees Celsius.

A similar seasonal pattern is found in standard error of measurement error. Figure 2.5 (b) shows how the standard error of measurement error varies across months. The solid line in orange shows coefficients on months for all states in the sample. Each point on the line represents how standard error of measurement error changes relative to the standard error in July. Measurement error reaches peak in January, and drops to a trough in summer. The lines in green and navy represents sets of coefficients for California and Texas respectively. Texas's standard error of measurement error varies more with season, and California's standard error of measurement error

varies less with season. In other words, measurement error's influence on climate impact models is more consistent in places like California than in places like Texas where seasonality is more salient.

2.4   Measurement Error in Climate Index

Besides directly depicting short term weather fluctuations, daily mean temperature can also be used to construct longer term measurement of weather variables. For example, heating/cooling/growing degree-days are functions of daily mean temperature, and are widely used measures of heating and cooling in energy and agriculture sectors. Binned regression model classifies daily mean temperature into varies temperature bins.

In section 2.4.1, I use heating and cooling degree-days as examples, and explore how presense of measurement error affects the construction of climate index. In section 2.4.2, I take weather derivative as an example and discuss how measurement error in climate index affects business risk management decisions.

2.4.1   Measurement Error in HDD and CDD

Heating degree day (HDD) is a measurement designed to quantify the demand for energy needed to heat a building. Likewise, cooling degree day (CDD) is a measurement designed to quantify the demand for air conditioning. Both HDD and CDD are defined relative to a base temperature, a comfortable temperature at which buildings need neither heating nor cooling. In the U.S, this base temperature is defined at $18°C$ ($65°F$).

Daily HDD is calculated as the greater of $18 - T_{mean}$ and zero, where T_{mean} is the daily mean temperature. Similarly, daily CDD is calculated as the greater of $T_{mean} - 18$ and zero. HDD

and CDD take non-negative values by definition. The monthly HDD and CDD are found by summation of the daily values for each location over a month.

Since daily mean temperature is the key input to calculate HDD and CDD, measurement error would be introduced in the construction of HDD and CDD. I calculate monthly HDD for October to April and monthly CDD for April to October using both true ($\tilde{T}_{mean}$) and mismeasured ($\hat{T}_{mean}$) daily mean temperature. I then construct measurement error in monthly HDD as Err.HDD= $\widehat{HDD} - \widetilde{HDD}$, and similarly measurement error in monthly CDD as Err.CDD= $\widehat{CDD} - \widetilde{CDD}$. Using these constructed measurement error, I characterize features of measurement error in monthly HDD and CDD.

1. Measurement error in HDD is larger than in CDD.

Table 2.3 provides summary statistics for measurement error in monthly HDD and CDD. The average of HDD is about three times the average of CDD, and the standard error of HDD is twice the standard error of CDD. This result carries to the magnitude and variance of measurement error: HDD bears larger measurement error and more noise in measurement error than CDD.

Table 2.3: Summary Statistics for Measurement Error in Monthly HDD and CDD

	Mean	SD	Min	Max
True HDD (°C)	356.34	(231.12)	0.00	1567.66
Mismeasured HDD (°C)	339.83	(230.54)	0.00	1565.05
Err. HDD (°C)	-16.51	(10.50)	-69.97	23.88
True CDD (°C)	99.15	(102.13)	0.00	582.42
Mismeasured CDD (°C)	107.04	(106.09)	0.00	588.75
Err. CDD (°C)	7.89	(8.26)	-58.67	56.04

2. HDD is underestimated and CDD is overestimated.

96

HDD tends to be underestimated and CDD tends to be overestimated as shown in Figure 2.6, which is a density plot of measurement error in HDD and CDD. This finding is consistent with the fact that coarse data mean estimator often overestimates daily mean temperature, so that when plugging in these overestimated daily mean temperature into threshold functions, the direction of error is mechanically determined. Details can be found in Appendix A.
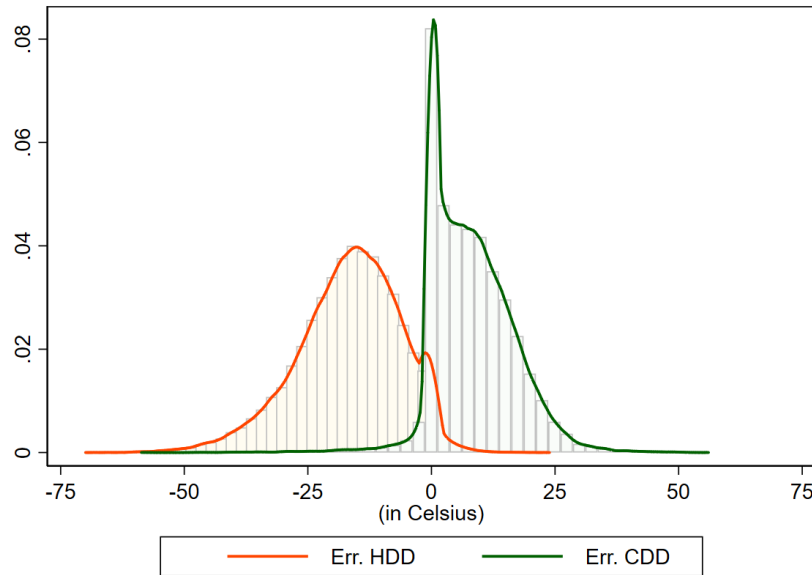

Figure 2.6: Density Plot of Measurement Error in Monthly HDD and CDD

Notes: Measurement error distributions. Orange line represents measurement error distribution in monthly HDD, and green line represents measurement error distribution in monthly CDD. Epanechnikov kernel density functions are used.

3. Measurement errors in HDD and CDD are nonclassical.

Mean zero assumption clearly would not hold based on the previous conclusion that HDD is often underestimated while CDD overestimated. Furthermore, from Figure 2.6 one can see that these compound measurement errors are not normally distributed.
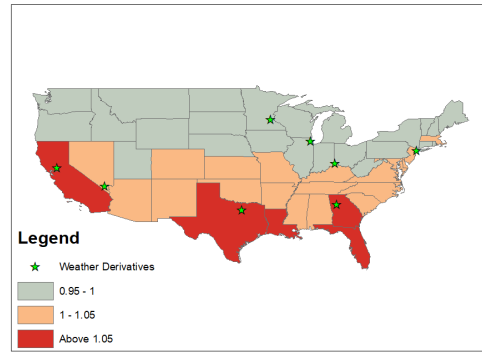
4. Measurement errors in HDD and CDD vary across location.

Figure 2.7 shows state average mean and standard error of measurement error in monthly HDD and CDD. One can see from Figure 2.7 (a) that HDD is largerly underestimated in mountain areas. In other words, using mismeasured HDD would underestimate amount of heating needed for buildings during winter in these places. In only a few states (e.g., FL, ME, VT, and MI), usinig mismeasured HDD would overestimate heating needed for buildings.

Figure 2.7 (c) is the plot of state average mean of measurement error in monthly CDD. During summer, coarse data mean estimator tends to overestimate CDD, especially for states in the South, which implies that buildings in these areas do not need as much heating as suggested by mismeasured CDD. The only exception is Nevada, which needs more cooling than mismeasured CDD suggests.
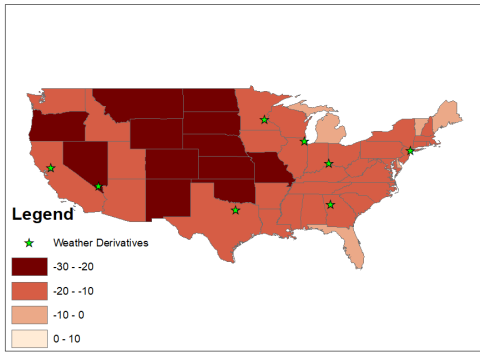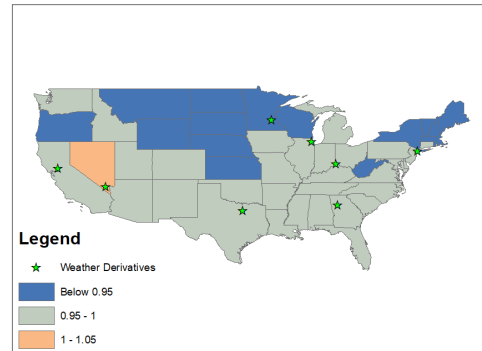
(a) Sample Mean of HDD Measurement Error

(b) Bias Factor in HDD
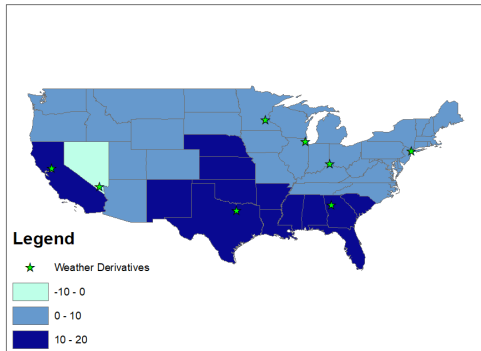
(c) Sample Mean of CDD Measurement Error

(d) Bias Factor in CDD

Figure 2.7: Measurement Error Statistics in Monthly HDD & CDD, by States

Note: These maps plot the spatial distribution of measurement errors in monthly HDD and CDD, averaged over the period from 1981 to 2010 for each state. Ranges in the legend are in terms of degrees Celsius. Subfigure (a) (c) plot sample means of measurement error within each state for HDD and CDD. Subfigure (b) (d) plot bias factors in HDD and CDD calculated from the degree-days version of $1 - \hat{b}_{u,T}$. No bias implies bias factor equals to 1. Eight cities listed in CME weather derivatives are labeled with green star.

Besides directly used as indexes for how much energy is needed to heat or cool buildings, HDD and CDD are also used widely in climate literature as explanatory variables. Similar to the measurement error framework described in section 3.2, it is the second moments of measurement error that is critical for evaluating the influence of measurement error in climate impact models. I calculate the fraction term in equation (2.2) for HDD and CDD in each state.

Figure 2.7 (b) and (d) show bias factors for HDD and CDD in each state. A clear location pattern exists for both HDD and CDD: when using HDD as explanatory variable, northern states suffer from attenuation bias and southern states suffer from overestimation problem. The magnitude of bias gets larger in degree-days statistics compared to in daily mean temperature. In southern states, bias factors in HDD increase by about 5% as compared to bias factors calculated in daily mean temperature in Figure 2.3 (b). When using CDD as explanatory variable, most states suffer from attenuation bias, except Nevada. In northern states, bias factors in CDD decrease by about 5% as compared to bias factors calculated in daily mean temperature in Figure 2.3 (d).

5. Measurement errors in HDD and CDD vary across season.

Measurement errors in HDD and CDD follow seasonal pattern as well. I examine how measurement errors in HDD and CDD vary across months. I regress level and standard error of measurement error in monthly HDD and CDD using October as the base level[3]. Standard error is calculated for each state.

Figure 2.8 shows how the magnitude and standard error of measurement error varies across months. The solid line in orange shows coefficients on months for measurement error in HDD. Measurement error in HDD is generally negative (as demonstrated in Figure 2.6 orange line), so that the lower the coefficient, the larger the measurement error. Therefore, the magnitude of measurement error increases as winter persists and returns to base level when winter ends. The dotted orange line depicts standard error in the measurement error of HDD, which is an indicator for measurement error's influence on climate impact estimation. Noise in HDD peaks in January and decreases afterwards.

---

[3] April and October are the only two months used for both HDD and CDD calculations.
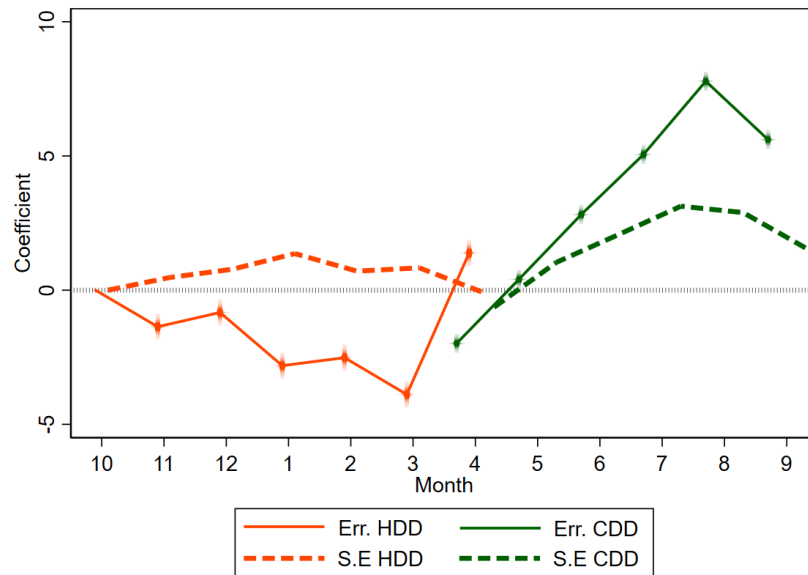
Figure 2.8: Seasonal Patterns of Measurement Error in HDD and CDD

Notes: This figure plots the temporal distribution of measurement errors in HDD and CDD. Coefficients in HDD are derived from regressing measurement errors and standard errors of measurement errors in monthly HDD on month dummies (Oct to next Apr), with October as the base level. Coefficients in CDD are derived on month dummies (Apr to Oct), also with October as the base level. Standard error of measurement error in monthly HDD and CDD are calculated for each station across year 1981 to 2010. All coefficients are in terms of degrees Celsius.

The solid line in green shows coefficients on months for measurement error in CDD. Measurement error in CDD is positive (as demonstrated in Figure 2.6 green line), so that the larger the coefficient, the larger the measurement error. Measurement error in CDD increases till August, then drops in September and return to base level in October (omited). The dotted green line depicts standard error in the measurement error of CDD. Noise in CDD peaks in July and decreases afterwards.

Overall, I show that measurement errors in HDD and CDD have more influence on extremely cold and hot months, and these months are also more interest from a policy perspective.

2.4.2   Cost of Measurement Error in Weather Derivative Market

It is estimated that nearly 30% of the U.S. economy is directly affected by the weather. Companies whose business depends on the weather, such as businesses in energy, sports management, agriculture, travel and others often need to manage risk associated with weather volatility. Weather derivatives are developed in the late 1990s to fill this growing need of weather risk management. The Chicago Mercantile Exchange (CME) lists climate index-based weather derivative products. To be specific, the price of weather derivatives traded in the winter and summer are based on monthly HDD and CDD values respectively, with 20$ per contract attached to each degree-day. For example, an utility in New York can experience a large decrease in revenue during mild summer due to low cooling demand. Each lost CDD during the summer decreases earnings by 50,000$. Therefore, this utility might need to purchase an option that pays 50,000$ per CDD below the pre-determined strike price to hedge against risk of cool summer.

Weather derivative creates a setting in which weather is quantified with an explicit price. Companies use statistical model to generate the possible range of outcomes that the underlying weather index may take, and lost associated with changes in HDD and CDD. This setting also enables calculation of cost associated with measurement error in the traded weather indexes. The discrepancy between different methods to construct climate index can potentially affect company's risk management decisions.

In this section, I quantify average cost of measurement error per contract associated with monthly HDD and CDD for all eight U.S cities listed. The temperature for a particular city is reported from an major airport weather station. The eight U.S cities and respective airport codes are listed in the first two columns of Table 2.4.

CME calculates HDD and CDD using $(\tau_{min} + \tau_{max})/2$ as its input for daily mean temper-

Table 2.4: Measurement Error Cost in Weather Derivatives

City	Airport Code	HDD(°F)	HDD($)	CDD(°F)	CDD($)
Atlanta	ATL	-28.46	569	17.40	348
Chicago	ORD	-24.56	491	9.25	185
Cincinnati	CVG	-26.74	535	11.38	228
Dallas	DFW	-31.96	639	19.63	393
Las Vegas	LAS	-23.49	470	4.70	94
Minneapolis	MSP	-26.02	520	10.63	213
New York City	LGA	-25.91	518	15.44	309
Sacramento	SAC	-41.32	826	50.86	1017

ature. This has created two potential measurement error problems in company's risk management decisions. The first measurement error issue is familiar to people: HDD and CDD measured with measurement error are used in regression analysis to study the impact of climate on economic outcome variables. Presense of measurement error leads to bias in the estimated impact as discussed in the previous sections, and companies can over- or underestimate impact of weather on their outcome variables of interest. As a result, they might over- or under-invest in weather and other derivatives. Figure 2.7 labels eight cities listed for trading. Estimates of HDD impact for cities in the South (e.g., Sacramento, Dallas, Atlanta) bear sizable overestimation bias. Estimates of CDD impact for most cities are subject to attenuation bias, and more severe for Minneapolis than elsewhere.

The second issue is that a company's cost-benefit analysis can be sensitive to methods in the construction of certain summary statistics. In a simplified setting, suppose there are three hours a day, and hourly temperature takes value of 0 and 1 only. Company incurs x$ loss per hour when hourly temperature is x. There are two temperature profiles A: [0 0 1] and B: [0 1 1]. When using coarse data estimator to estimate daily mean temperature, both profile has daily mean temperature of 0.5, which translates to a loss incured during the day of 1.5$. When using hourly temperature to estimate loss, profile A leads to 1$ and profile B leads to 2$ loss.

Therefore, estimates of average revenue and cost on weather is asssociated with how weather index is constructed. A bad estimate of cost associated with weather risk can lead to over- or under-invest in weather and other derivatives.
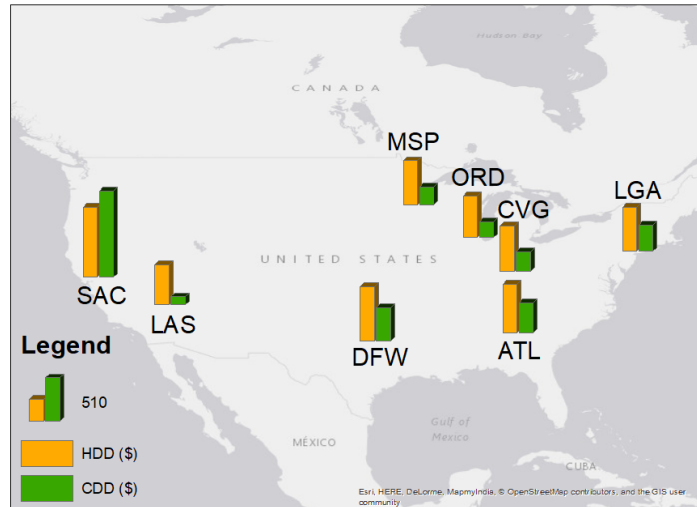

Figure 2.9: Average Cost of Measurement Error per Contract

Notes: This map plots first order average cost of measurement errors per contract in monthly HDD and CDD for eight cities listed in CME weather derivatives. CME uses temperature data collected from corresponding major airport stations. The green bar in legend is equivalent to 510$.

I calculate average cost of measurement error per contract for both monthly HDD and CDD. The average measurement error in monthly HDD and CDD, as well as their monetized value per contract are reported in Table 2.4. Average cost in measurement error of monthly HDD is about 570$ per contract in these cities, and this cost is 345$ per contract for monthly CDD. Figure 2.9 is visualization of the monetized cost of measurement error per contract for all eight U.S cities listed. Cost of measurement error is larger for HDD than for CDD in all eight cities except Sacramento. Sacramento is also the place where cost of measurement error is the largest for both HDD and CDD, and this cost is above 1,000$ per contract.

2.5   Conclusion

This paper explores the potential role of measurement error induced by using coarse data to approximate daily mean temperature. I find substantive differences between daily mean temperature constructed using the traditional (average of Tmax and Tmin) and 24 hourly observations (average of hourly temperature) by constructing measurement error using various sets of U.S. weather data product spanning 1981 to 2010. This paper also extends to measurement error in two climate indexes, heating and cooling degree-days, which take daily mean temperature as key input. I report three main findings. First, measurement errors in daily mean temperature and related climate indexes are nonclassical in general, and does not always lead to attenuation bias. Second, measurement errors vary significantly across locations and seasons. Finally, the impact of measurement error in temperature variables is ambiguous. States in the south are more prone to overestimation while states in the north are more prone to attenuation bias. The magnitude of bias increases in cumulative temperature measures like degree-days.

Overall, this work has strong policy implications. My result shows that measurement error induced by using coarse averaging method can lead to around 10% of bias in either direction in the climate impact estimation. This number is rather conservative for two reasons. First, measurement error characterized in this paper is only one of the four sources of measurement error summarized in chapter one. I already show that using PRISM reanalysis data to construct mean temperature combines two sources of measurement error, and incurs larger noise in climate impact estimation than using ground weather station data. In fact, existing literature typically bear two to three types of measurement error, thus influence on climate impact models is larger than what is estimated in this paper. Second, there are microclimate conditions not explored in this study. The size and impact of measurement error could vary widely depending on the scale of the study.

2.6   Acknowledgements

Chapter 2 is being prepared for publication. The dissertation author is the sole researcher on this chapter.

2.7   Additional Figures and Tables

2.7.1   Measurement Error in HDD and CDD

Measurement error induced by using coarse data is usually negative in HDD and positive in CDD. This follows mechanically from the fact that measurement error in daily mean temperature is found to be positive ($\hat{T}_{mean} - \tilde{T}_{mean} > 0$).

Suppose $u = \hat{T}_{mean} - \tilde{T}_{mean} > 0$, then:

$$\text{True HDD} = max(18 - \tilde{T}_{mean}, 0)$$

$$\text{Mismeasured HDD} = max(18 - \hat{T}_{mean}, 0)$$

$$= max(18 - u - \tilde{T}_{mean}, 0)$$

Since $u > 0$ by construction, True HDD > Mismeasured HDD is always true because of the function form of HDD. I plot True HDD (in green) and Mismeasured HDD (in orange) in Figure2.10 (a). Measurement error in HDD is negative when true daily mean temperature is below 18°C. When summing across daily HDD to obtain monthly HDD, this negative measurement error is amplified. This explains why monthly HDD is underestimated in section 2.4.1.

Similarly,

$$\text{True CDD} = max(\tilde{T}_{mean} - 18, 0)$$

$$\text{Mismeasured CDD} = max(\hat{T}_{mean} - 18, 0)$$

$$= max(\tilde{T}_{mean} - (18 - u), 0)$$

True CDD $<$ Mismeasured CDD always holds because of the function form of CDD. I plot True CDD (in green) and Mismeasured CDD (in orange) in Figure2.10 (b). Measurement error in CDD is positive when true daily mean temperature is below 18-u °C. When summing across daily CDD to obtain monthly CDD, this positive measurement error is amplified.
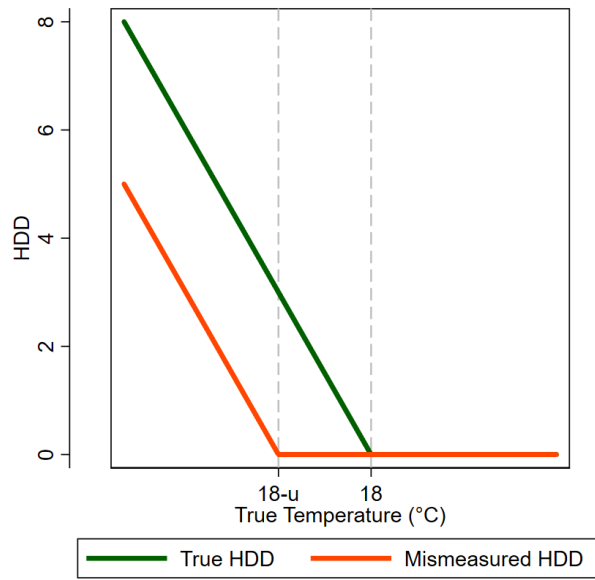
(a) HDD Function
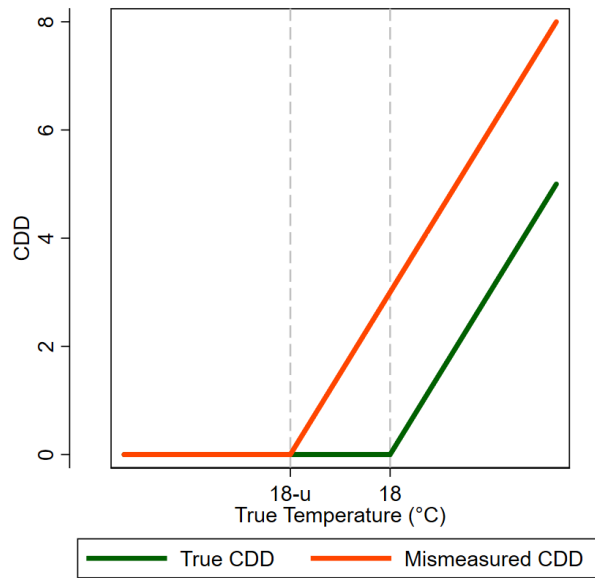
(b) CDD Function

Figure 2.10: Measurement Error in HDD and CDD

Chapter 3

Estimating the Impact of Climate Change: An Exploration of the Bin Regression Model

3.1   Motivation

In recent years, a large number of empirical works have been undertaken to study the influence of climate variables on a wide range of different social and economic outcomes. As noted by [Hsiang, 2016], the measurement and representation of these climate variables is a critical step in identifying the impacts of climate change. The "bin" regression model, which is a flexible semi-parametric method for representing one or more of these climate variables, has emerged as the workhorse approach for empirical work (e.g., [Auffhammer and Aroonruengsawat, 2011]; [Deschênes and Greenstone, 2011]; [Graff Zivin and Neidell, 2014]; [Barreca et al., 2016]).

This paper is the first to formally explore the econometric properties of the bin regression approach in climate economics literature. The bin regression approach takes the desired explana-

tory variable and discretizes in a manner like a histogram with each bin now being represented by a count (e.g., number of days during the growing seasons where mean temperature falls in a specified interval). Formally, the bin approach is a M-piecewise constant function, where the intervals defining the M bins are chosen by the researcher. A core implicit assumption with the bin approach is that the distribution of the climate/weather variable is not adequately represented by a data generating process (DGP) where a single sufficient statistic such as the growing season's mean is not adequate, with the usual rationale being the belief that non-linearity in the response is important in one or both of the tails. The bin approach also implicitly assumes that the temporal ordering of individual observations (e.g., daily means) is not important as this detail is lost when the explanatory variable is binned.

We show that, although the bin regression approach often produces reasonable results, it produces consistent parameter estimates only under very stringent and unlikely assumptions on true data generating procedure, which we characterize in detail. Furthermore, because the researcher chooses the bin definitions, the approach is not truly semiparametric. It is possible to rectifying this issue by estimating the number of bins and their defining thresholds, but this dramatically increases what is already a non-parsimonious specification since researchers typically use ten or more bins.

The bin approach has two other major problems. First, because information within a bin is lost and the bins themselves represent points of discontinuity, it is possible for an overall shift in the distribution of the variable of interest such as an average increase in temperature either have little or an extreme effect. A small change is indicated if the original binning tended to place observations away from the right-defining points of bins so that few observations shift between bins. Conversely, an extreme effect may be indicated if the original bin definitions tended to concentrate observations near the right defining-points so a sizeable fraction of observations is

predicted to jump from one bin to a higher one. Second, we show that measurement error [Carroll et al., 1995] in the original variable of interest interacts with the bins in non-neutral ways that can produce clear signs of a non-linear response function when that is not the case.

We propose alternatives to bin model for estimating the impact of climate change that assume the same underlying DGP. The main line of attack we take comes from recognizing that the bin regression approach conflates two problems: (a) summarize the distribution of an exogenous stimulus variable and (b) determine how the response variable of interest changes as the distribution of the stimulus variable changes.

Several standard approaches to summarizing the distribution of the stimulus variable of interest are considered ranging from kernel density estimation to fitting mixtures of normals. Effectively, the question is how many moments/summary statistics of the stimulus variable are needed to adequately represent the unknown distribution, which is a well-researched area of statistics. While it would be unusual to need more moments/summary statistics to represent the stimulus variable than the number of bin parameters, it may not be immediately clear how these should influence the response variable. To this end, we examine suitability of modern variable selection methods (e.g., MARS) that control for overfitting to determine which indicators of the stimulus variable distribution and their appropriate transformation are needed for a parsimonious representation. This potentially allows for summary statistics like quantiles of the stimulus variable distribution, which are generally functions of moments/summary statistics to be included as potential predictors. It also allows for the inclusion of interactions between these stimulus variable statistics and other covariates in the variable selection process, which facilitates an informative representation of whether the response to changes in the stimulus variable is heterogenous with respect to the observables.

The properties of the bin model and alternatives are further illustrated with monte carlo simulations and empirical datasets. We explore potential alternatives with better statistical properties. One direction is to pursue more flexible binning, as the methods used in [Schlenker and Roberts, 2009]. We show in the paper that their method are more flexible than exogenously defined binned regression model. In addition, we show that one can also incorporate the weather distribution parameters into regression models, particularly if the main purpose is to look at a smooth shift in the temperature distribution either due to measurement error or for climate damage projection purpose.

The rest of the paper is organized as follows. Section 3.2 discusses a set of issues related to the use of bin regression model. Section 3.3 explores potential alternatives to the bin model that could potentially overcome issues mentioned in section 3.2. Section 3.4 concludes this paper.

3.2 Issues with the Binning Approach

3.2.1 Specification of Binned Regression Model

As noted in [Dell et al., 2014], a standard panel approach to estimating the relationship between climate and outcome variables in this literature typically takes variations of the form

$$y_{it} = f(W_{it}) + \gamma Z_{it} + c_i + \theta_t + \varepsilon_{it} \tag{3.1}$$

where y_{it} is the value of the dependent variable of interest, such as mortality, GDP growth, test score etc. Z_{it} is a vector of other covariates. μ_i and θ_t are fixed effects for units and time. ε_{it} is a well-behaved error term.

$f(\cdot)$ is an unknown functional form to be approximated using "bins" and the weather variable W_{it}. Suppose temperature is the only climate variable in the above specification, and define $j = 1, ..., J$ bin and count the number of days at location i in year t in each bin as $\text{TBIN}_{j,it}$. We may rewrite equation (3.1) as:

$$y_{it} = \sum_{j=2}^{J} \text{TBIN}_{j,it} + \gamma Z_{it} + c_i + \theta_t + \varepsilon_{it} \tag{3.2}$$

where the first bin $j = 1$ is omitted to avoid multicollinearity.

All such bin regression models take observation of W_{it} and place it into a pre-defined bin and use the information about bin membership to predict an outcome variable. The difference across these models are on whether outcome variable is observed at a lower frequency than the weather variable. Different ways of defining bins result in binned regression models with different statistical properties. Whether the weather variable is observed at the same or higher frequency than the outcome variable affects the function of the bins in the model.

1. **Same frequency.** (e.g., daily outcome and temperature). For each individual weather observations, bins are indicator variables. In such cases, each temperature observation is translated into 10 indicator variables, which equals 1 if an observation falls into that temperature bin and 0 otherwise. For example, in [Park, 2018], both the outcome variable test score, and maximum temperature is observed at daily level.

2. **Higher frequency.** (e.g., annual outcome and daily temperature). For each individual weather observations, bins are count variables. In such cases, each temperature observations are categorized into one of the 10 bins, and then the number of days in each temperature bin is counted. For example, [Deschênes and Greenstone, 2011] observe outcome variable mortality rate at annual level while temperature and precipitation variables at daily level.

Essentially, binning is a way to summarize weather distribution, and the bin regression model takes the outcome variable as a function of the weather distribution parameters.

In the case of same frequency outcome variable, there is a loss of information from binning. Suppose we exogenously fix number of bins and the bin width, and for simplicity assume there is no covariate. The true DGP for y_{it} is that $E[y_{it}|W_{it}] = f(W_{it})$, where $f(\cdot)$ is a smooth function with continuous derivatives. Using bins, the best case scenario assumes that $E[y_{it}|W_{it} \in Bin_k] = f(\bar{W}_k)$, where $\bar{W}_k$ is the mean of the W_{it} in bin k that contains the particular W_{it} of interest. As the divergence of $f(W_{it})$ from $f(\bar{W}_k)$ increases, the error associated with binning also increases.
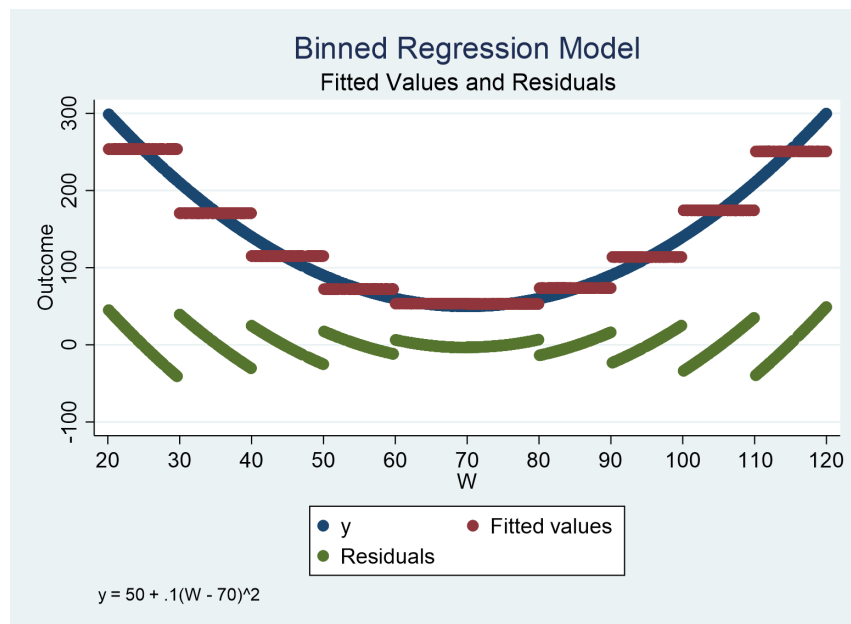

Figure 3.1: Loss of Information from Binning

Let us define the true DGP as $y = 50 + .1(W - 70)^2$. We then fit y with a typical 10 bin regression model. Figure 3.1 plots the fitted values and residuals of the bin regression. Following the convention in the literature, this bin regression model has unlimited bounds at the lowest and highest bins (i.e., the lowest bin is below 30°F and the highest bin is above 110°F). The quadratic

curve in navy represents the true DGP. There are two key observations from this simple example: first, the prediction is always constant within a bin even though the function was changing; second, the residual gets larger as W moves away from the mean or median temperture of the bin, and this effect is more pronounced for the tail bins than for the middle ones.

In the case of higher frequency weather variables, there are three different scenarios. First, if the outcome Y_{it} is a linear function of first moment of the weather variables, then bin model is grossly inefficient, as we can simply model the climate's impact with an annual mean of weather variables.

Second, if the outcome Y_{it} is a function of the probability density function (pdf) of $W_{it} = \{w_{it1}, ..., w_{itD}\}$, then one can think of bin regression model as a way to summarize W_{it}, which might not be the optimal summary statistics as we will illustrate more in the following subsections.

Third, if the outcome Y_{it} is a distributed lag model of daily weather variable w_{itd}, where the order of these daily observations matters, then bin regression models are likely going to miss key DGP features. For example, consecutive hot days kill more people; a heavy rain spell can either enhance crop product or destroy it depending on when the spell occurs. Furthermore, there are also interaction effects between weather and other variables which are not captured in these bin models.

3.2.2 Defining Bins

In the case of high-frequency weather variables, bins can be thought of as a variant of a histogram representation of pdf. The properties of different types of histograms are explored in depth in statistics literature (e.g., [Scott, 2015]). There are three key features associated with any

bin definition: number of bins; bin boundaries/endpoints; and bin widths-same for all bins or not.

In the bin literature, typically bins are determined using different rules. First, bins can be exogenously defined independent of data. Second, bins can determined looking at W_{it} using simple exogenous rule, and typical climate economics literature tends to apply this rule. For example, one can split the pdf from left to right into 10 equal groups so that each bin contains the same number of observations. Third, bins can be chosen using data to optimize some objectives. For example, one can minimize sum of squared residuals subject to certain constraints.

Typical Setup in Environmental Economics

We focus on the case where weather variables are observed at a higher frequency. For example, we observe daily temperature $\{w_{it1},...w_{itD}\}$ and annual outcome Y_{it}. In the typical setup in environmental economics literature, the number of bins are fixed exogenously and people typically use 10 bins. There are usually 8 bins in the middle with equal bin width, and unlimited bounds for the lowest and highest bins, thus end bins have bin widths that are different from the middle bins, and usually much larger bin width.

The bin regression is usually described as a semiparametric regression, however, it is not necessarily a correct characterization. We compare three different sets of bin definition, and show that quality of approximation varies with how bins are defined. Given known sample size and desired number of bins, we compare three bin regressions with the following bin definitions:

- Fix number of observations in each bin

- Fix bin widths

- Equal bin widths in the middle, unlimited bounds for end bins

The temperature data used in this exercise come from daily mean temperature spanning 1960-2017 for more than 3,000 weather monitors in the U.S. The true relationship between outcome and temperature is quadratic as shown in the top left panel in figure 3.2. Each point on the blue line can be interpreted as " the increase in Y associated with one additional day at temperature X", which is the same as the intepretation for bin coefficient. Details on the simulation procedure can be found in section 3.6.1.
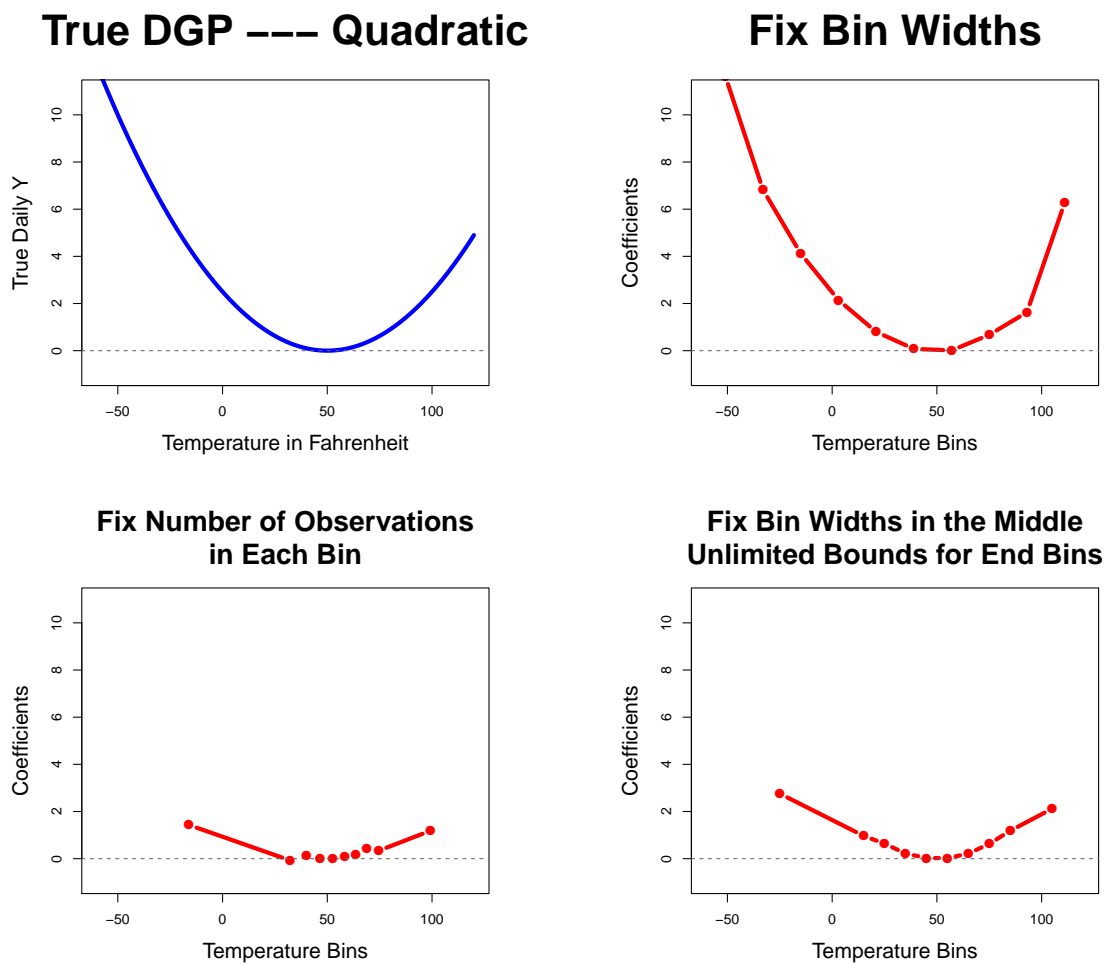

Figure 3.2: Quality of Approximation Depends on Bin Definitions

When bin widths are fixed so that bins have equal width, the bin coefficients line up nicely

and the shape of the response function mimics that of the true DGP as shown in top right panel in the graph. When number of observations in each bin is fixed, middle bins are narrow and end bins cover a wide range of temperature observations. As a result, impacts at the end bins are not well modeled because the distribution of the temperature variables at the tails are dispersed, so that estimating the highest temperature bin is essentially estimating the temperature effect at 100°F, assuming temperature at 90°F and 110°F have the same effect on outcome as temperature at 100°F. Fixing bin widths in the middle and leaving unlimited bounds for end bins is the common practice in bin regression model, where a similar problem at the tail bins occur. In other words, the typical bin regression model used in the literature could hardly identify the impact on outcome for moving temperature observations from 90°F to 100°F.

The take-away message is that the quality of approximation varies with how bins are defined. In fact, refering to the histogram of daily temperature observations, the divergence in approximation can be easily understood. Figure 3.3 shows a histogram of daily temperature used in the simulation exercise. Green bars represent the first decile and purple bars represent the last decile. It is worth noticing that the first decile covers a range of 80°F and the last decile 40°F, while all other deciles cover less than 10°F. The end bins are covering a wide range of temperature observations so estimating the ever-changing tail bin effects with a single bin estimator is almost never a good practice.
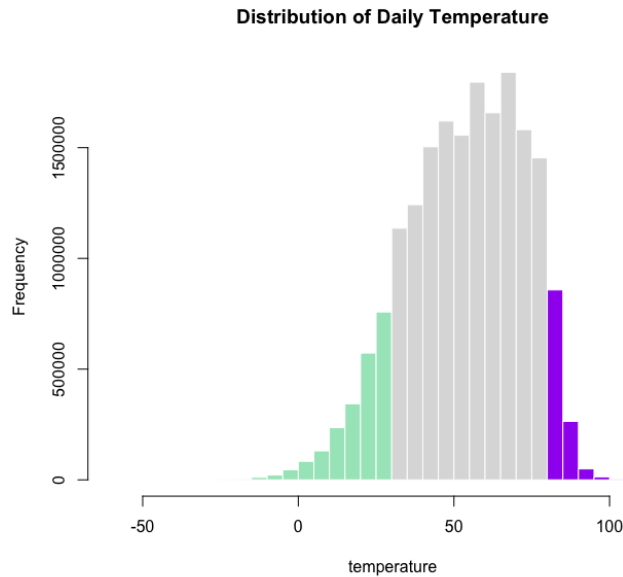
Figure 3.3: A Histogram of Daily Temperature
Note: The green bars denote first decile, and purple bars denote last decile.

Nature of Bin Regression: A Flexible Functional Form

The nature of the bin regression is essentially a step function and a special case of a local constant regression model. It has the same problem as other step function estimators like the Turnbull estimators used in contingent valuation. In both cases, it is relatively difficult to pin down the tail effect. The closest analogy to bin regression model is a regression spline model. Here are some variants of such regression spline models in order of increasing flexbility.

1. Fixed number of knots; fixed knot locations; fixed slopes between knots

   - Bin regression model is an example of this stringent variant in which the number of knots are fixed (usually 11); knot locations are pre-defined for every 10°F; zero slope between any two consecutive knots.

2. Fixed number of knots; fixed slopes between knots; free to pick knot locations

3. Fixed number of knots; free to estimate slopes between knots; free to pick knot locations

- Piece-wise linear model is one such example in which given the number of knots, one can pick the location of each knot and estimate the slopes between knots.

4. Free to pick number of knots; free to estimate slopes between knots; free to pick knot locations

- This procedure allows the number of knots to increase as sample size increases and allows slopes between knots to be estimated. For example, one can use generalized cross validation to choose number of knots and other tunning parameters.

3.2.3 Interaction of Bins with Measurement Error

Measurement of weather variables is a critical issue for identifying impacts of climate. An often documented, but rarely dealt with, problem is the deviation between the actual value of the climate variable of interest and the proxy variable used in the estimated climate impact models ([Auffhammer et al., 2013]). This deviation is known as measurement error in weather variables, and is recently characterized by [Carson and Yu, 2020]. The key finding is that measurement errors in temperature statistics come from multiple sources, and are non-classical and often large in noise/signal sense.

Measurement error in weather variables coupled with binning often leads to misclassification error, which is an obvious consequence. Perhaps what is less obvious and more harmful is that these misclassification errors are highly asymmetric: it is the end bins that are more prone to misclassiciation error, yet they are also the bins of most interest from a policy perspective.

We demonstrate how measurement error interacts with binned regression model with a simulation exercise. We generate temperature observations T_0 following a $N(50, 16^2)$ distribution, then add i.i.d.$N(0, 1)$ error to the temperature observations to generate a new set of mismeasured

temperature observations T_1. The correlation between T_0 and T_1 is about .85. In other words, the measurement error is about 40% of the random variation in true temperature:

$$T_0 = 50 + 16 \times N(0,1)$$

$$T_1 = T + 10 \times N(0,1)$$

We create 10 bins for both T_0 and T_1 in 10°F bins. Figure 3.4 displays what percentage of each type of observations shows up in each of the temperature bins defined on the temperature with measurement error. Navy bars denote percent of observations that come from a lower true bin, gray bars denote observations match in the sense of being in the same bin while red bars denote an observation that come from a higher bin. Note that the lowest bin can not have any "lower true bin" observations by construction while the same is true of "higher true bin" for the highest bin.
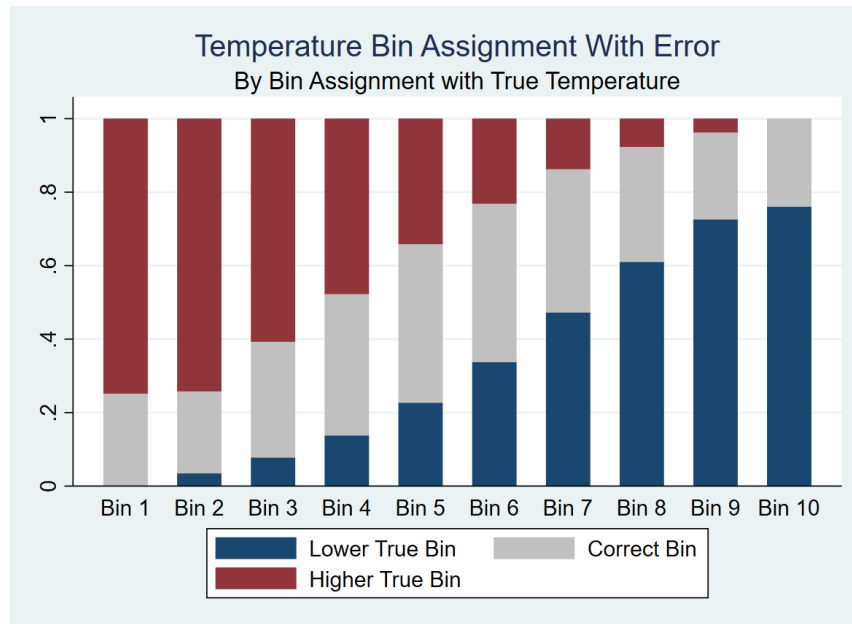

Figure 3.4: Temperature Bin Assignment w/ Error

There are two striking observations: First, a very large fraction of the observations are

missclassified in the sense that T^* and T would appear in different bins. Second, there is pronounced asymmetry in the number of misclassified observations in each bin, except in the middle of the distribution, with this asymmetry changing in a predictable way from small to large temperature bins. Less obvious, but perhaps worth noting, the means/medians of the lowest and highest bins are different (pulled toward the mean) relative to the means/medians in the true bins.

3.2.4   Role of Bins in Forecasting Impact of Climate Change

Bins can be either insensitive or hyper-sensitive to predicted changes in temperature distribution. This issue is more prounounced for the end bins because the end bins are often open-ended, thus a shift in distribution may not be captured and identified under binned regression model. For example, Table 3.1 lists three series of temperature observations. The first row represents the baseline temperature; second and third rows are two climate projections in which temperature increases in one of the three days. In projection 1, temperature in day 1 increases by $9°$F. However, if we look at the histogram representation on Figure 3.5, the increase in temperature for day 1 is not captured in the binned specification.

Under projection 2, temperature on day 2 increases by only $1°$F. Now that the day 2 temperature hits the bin threshold or cutoff point, the histogram representation changes drastically in comparison to the baseline scenario as shown in Figure 1.2, even though the actual increase in temperature is not as much as that in projection 1.

Table 3.1: Temperature Under Different Scenarios

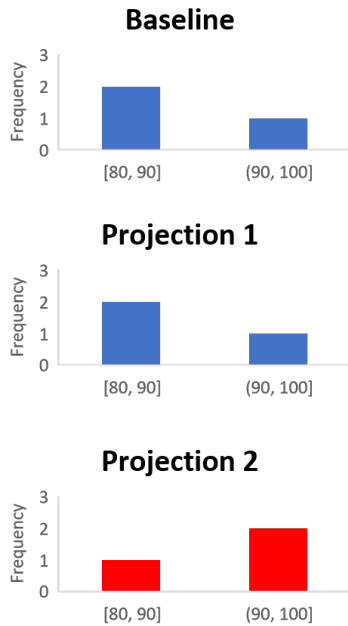|	Day 1	Day 2	Day 3
Baseline	81	90	91
Projection 1	90	90	91
Projection 2	81	91	91

Figure 3.5: Histogram of Temperature

3.2.5 Last Bin Accumulates Problems

The last bin is the one that accumulate all of the issues mentioned in the previous sections. However, this last bin is also the main object of interest for climate change, as people are often interested in knowing the effect of extreme heat or potential climate change impact. In typical economics application, last bin is usually open-ended, which makes several bin issues more pronounced: first, observations in the last bin is often sparsely populated, with potentially many outliers; second, observations within the last bin often concentrated near lower bin boundary. This is different from from the middle bins, where temperature observations are uniformly distributed

within the bin; third, the last bin covers a wide range of temperature yet the within bin respon-siveness is assumed constant, indicating that one can never project "hot days become hotter". In other words, we can never identify the effect of temperature moving from 90°F to 100°F; lastly, measurement error causes large and asymmetric misclassification, which typically will bias the tail bin estimated downward.

3.3 Alternatives to the Binned Regression Model

In this section, we explore some alternatives to the binned regression model under the assumption that knowing temperature distribution is sufficient for estimating the climate impact models. In fact, under this assumption, the modeling of the climate impact is essentially a two step problem: in the first step, estimate the empirical probability density function (pdf) of daily temperature, and obtain a group of distribution parameters; in the second step, estimate the outcome variable y as a function of the distribution parameters. The binned regression model combines these two procedures by estimating pdf as a histogram of temperature distribution, and setting the outcome variable as a linear function of the binned variables.

We begin this section by examing the properties of existing nonparametric and semi-parametric approaches proposed in the literature, which still use some variant of histogram representation to summarize time distribution of temperature, but apply a more generalized approach to the modelling the relationship between outcome and the temperature distribution parameters, where the bin model is a special case. In addition, we also explore some alternative representations of the temperature distribution other than histogram. For example, one can use some mixture normal models to approximate the empirical pdf, then model outcome variables as functions of the distribution parameters with various models.

As discussed in section 3.2.5, it is the last bin in the bin regression model that accumulates various problems, yet it is also the last bin that is of interest from a policy perspective. Therefore, one should expect the bin regression model to be a well-behaved model except at the tails. The goal for finding an alternative to the bin model is not necessarily finding an alternative model that could beat bin regression model everywhere, but to find an alternative that are more appropriate for approximating effect of extreme temperatures.

3.3.1 Histogram Representation of Temperature Distribution

We generate a temperature response model where daily outcome is a function of daily temperature, aggregate the outcome to annual level, and compare how different histogram representation-based models can be used to model the temperature response relationship.

Suppose the true response y_{id} for location i on day d is a function of the temperature on that day:

$$y_{id} = f(T_{id})$$

Let $f()$ take a quadratic form: $y = (T - 10)^2/1000$. We aggregate y_{id} by year and location such that $Y_{it} = \sum y_{id}, d \in t$ and bin the temperature observations. The final sample contains 2204 station-year observations, and is randomly divided into training and test samples, where the test sample is about 25% of the total sample.

We compare the performance of the models used in [Schlenker and Roberts, 2009] (SR), in which they estimate how crop yields respond to the time distribution of heat. Their key assumption is that yield growth $g(T)$ depends nonlinearly on heat T so that the log yield, y_{it} in location i and yeat t can be written as:

$$Y_{it} = \int_{\underline{T}}^{\overline{T}} g(T)\phi_{it}(T)dT^* + \delta z_{it} + c_i + \varepsilon_{it} \tag{3.3}$$

where $\phi_{it}(T)$ is the time distribution of heat over growing season in location i and year t. The observed temperatures T ranges between $\underline{T}$ and $\overline{T}$. Other covariates are denoted z_{it} and a location fixed effect c_i is used to control for time-invariant heterogeneity. For simulation purpose, we take the covariates and fixed effect out of the model. SR use a discretized representation of equation (3.3):

$$\int_{\underline{T}}^{\overline{T}} g(T)\phi_{it}(T)dT^* \approx \sum_{T=\underline{T}}^{\overline{T}} g(T + \frac{\Delta T}{2})[\Phi_{it}(T + \Delta T) - \Phi_{it}(T)] \tag{3.4}$$

where Φ is the cumulative distribution of temperature. Apply basis expansion on the growth function g:

$$g(T) = \sum_{k=1}^{K} \alpha_k \psi_k(T) \tag{3.5}$$

here $\{\psi_k\}_{k=1}^{K}$ is some basis such as polynomial or fourier basis. Combine equation (3.4) and (3.5):

$$\int_{\underline{T}}^{\overline{T}} g(T)\phi_{it}(T)dT \approx \sum_{k=1}^{K} \alpha_k \sum_{T=\underline{T}}^{\overline{T}} \psi_k(T + \frac{\Delta T}{2})[\Phi_{it}(T + \Delta T) - \Phi_{it}(T)]$$

$$= \sum_{k=1}^{K} \alpha_k x_{k,it} \tag{3.6}$$

where $x_{k,it} \equiv \sum_{T=\underline{T}}^{\overline{T}} \psi_k(T + \frac{\Delta T}{2})[\Phi_{it}(T + \Delta T) - \Phi_{it}(T)]$. Note that the pdf $[\Phi_{it}(T + \Delta T) - \Phi_{it}(T)]$ can be reasonably estimated for fixed bin width ΔT.

SR's paper explore three models based on this setting. In this section, we explore the statistics properties of these models with numerical exercises.

- **Binned regression model**. The binned regression model is similar in nature to the case where $\psi_k = \mathbb{1}_{[T+\frac{\Delta T}{2} \in \text{bin k}]}$. ΔT is the bin width, and $T + \frac{\Delta T}{2}$ is the midpoint of each interval. So that each $x_{k,it}$ represents the percent of days in temperature bin k for year t and location i. The model becomes:

$$Y_{it} = \sum_{k=2}^{K} {}^{1}\alpha_k \underbrace{\mathbb{1}_{[T+\frac{\Delta T}{2} \in \text{bin k}]} [\Phi_{it}(T+\Delta T) - \Phi_{it}(T)]}_{x_{k,it}} + \varepsilon_{it}$$

multiplying $x_{k,it}$ by 365 yields exactly the binned temperature variables typically used in the literature: number of days in each temperature bin.

Figure 3.6 plots the climate response function estimated with binned regression model. The solid line represents the true data generating process or how daily outcome variable y changes in respond to temperature. We allow bin width to vary and fit the binned regression model with bin width equal 1,3 and 5°C respectively. As is shown in figure 3.6, bin models deviates from the true DGP at the tail bins, because there are not enough observations at the end bins. From a policy perspective, it is often the tail bins that are of interest to empirical researchers, so even though the bin regression seems to lead to good fit in general, it could lead to wrong policy implications.
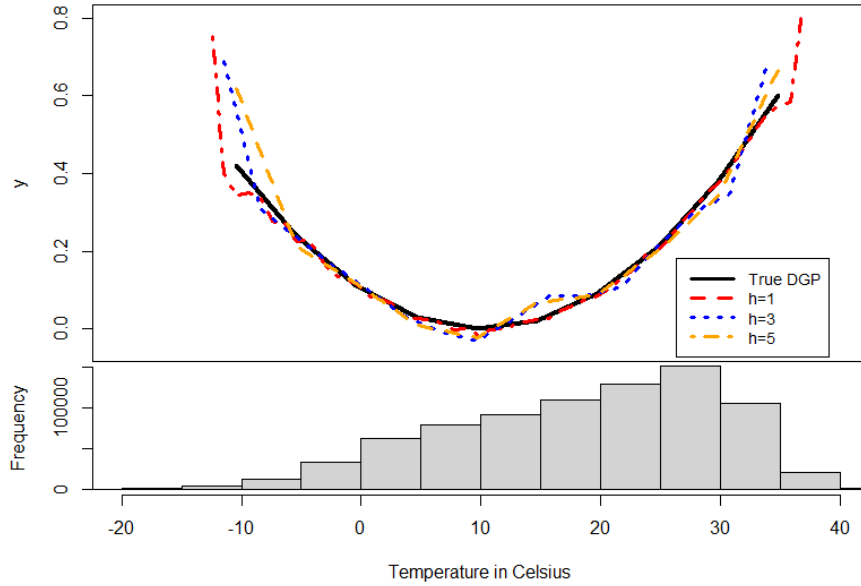
Figure 3.6: Binned Regression Model w/ different bin width

The choice of bin width also affects the goodness of fit. $h = 3$ (in blue, i.e, $\Delta T = 3°$) is the bin width used in SR's estimation and $h = 5$ (in orange) is the bin width used in most papers that adopt celsius temperature scale. $h = 1$ (in red) aligns the best with the true DGP. As bin width increases, the bin regression model deviates more from the actual quadratic curve. There are sizable deviations from the true DGP when temperature is $15°C$ and above $31°C$.

- **K-th order Chebyshev polynomial**. SR use a K-th order Chebyshev polynomial of the form $g(T) = \sum_{k=1}^{K} \alpha_k T_k(T)$, where $T_k()$ is the k-th order Chebyshev polynomial. Chebyshev polynomial is similar to Fourier basis, but the usage of Fourier expansion is limited for the circumstances where the estimated function are periodic.

$$Y_{it} = \sum_{k=1}^{K} \alpha_k \underbrace{T_k(T + \frac{\Delta T}{2})[\Phi_{it}(T + \Delta T) - \Phi_{it}(T)]}_{x_{k,it}} + \varepsilon_{it}$$

128

In the simulation exercise, we construct a set of 8-th order Chebyshev polynomial basis following SR's approach and let $\Delta T = 1°$. The curve for the Chebyshev polynomial model roughly lines up with the true DGP when temperature is above 18°C. The model does not perform well when temperature is at the low end, or when there are not enough observations within each 1°C bin. This problem with the use of Chebyshev polynomial is more severe when an alternative DGP is used. In section 3.6.2, we display the fit for Chebyshev polynomial model when DGP is exponential. As shown in figure 3.13, the polynomial fit oscilates around the true DGP, suggesting at least the polynomial model is sensitive to the underlying DGP.
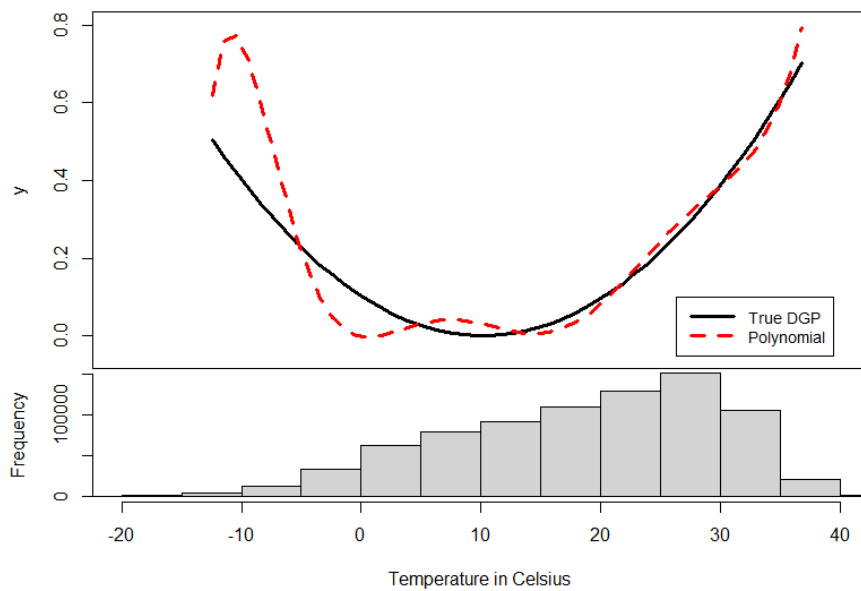

Figure 3.7: Chebyshev Polynomial Model

- **Piecewise linear model**. In the piecewise linear model, SR assume $g(T)$ to be a piecewise linear function, that allows outcome to linearly increase up to an endogenous threshold and then decrease linearly above the threshold. A typical piecewise linear function $g(T)$ can be written as:

$$g(T) = \begin{cases} -0.01T + 0.15 & \text{if } T < 15 \\[2ex] 0.03T - 0.57 & \text{if } 15 \leq T \end{cases}$$

In the simulation exercise, we fit a piecewise linear model with two intervals, although in practice, there are often more than one cutoff point. We loop over every interger threshold, and estimate the least-squares segment slopes and constant for each one. The model selection criterion used for cutoff selection is mean squared error. As shown in figure 3.8, the threshold temperature is 15°C, but the true cutoff point in the quadratic DGP is 10°C. Note that the use of piecewise linear model is not appropriate if one knows a quadratic DGP is likely. However, if the true DGP is a threshold model or a linear model, piecewise linear model can have much better performance. As shown in figure 3.14, piecewise linear model can nicely trace out the exponetial DGP when the polynomial model largely fails.
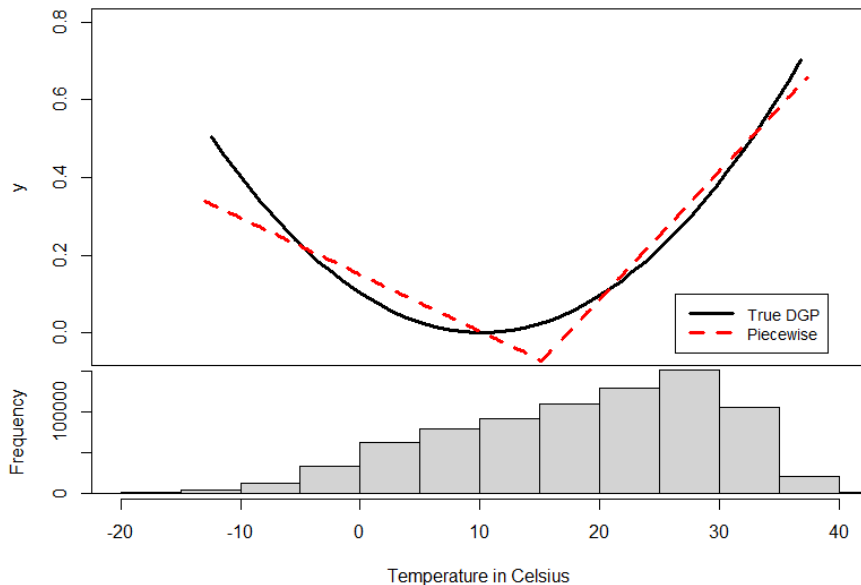

Figure 3.8: Piecewise Linear Model

3.3.2 Alternative Representations of Temperature Distribution

A two step estimator

In addition to representing the time distribution of temperature as histograms, we also explore other possible representations of temperature, which are potentially more robust to measurement error and more adaptive to projection problems. In particular, we propose a two step estimator under the assumption that temperature distribution can be estimated.

Assume T_{id} come from some arbitrary two parameter location-scale distribution (we take the example of a simple normal distribution for simplicity). Two steps are involved in order to estimate the temperature response function:

1. Estimate location and scale parameters of daily temperature T_{id}: μ_i and σ_i for each i

$$T_{id} \sim N(\mu_i, \sigma_i^2)$$

2. Estimate relationship between Y_i and the estimated location and scale parameters μ_i and σ_i

$$Y_i = g(\hat{\mu}_i) + h(\hat{\sigma}_i)) + k(\hat{\mu}_i, \hat{\sigma}_i) + \theta_i + \varepsilon_i$$

In practice, we can use more flexible distributions to approximate the time distribution of temperature. Figure 3.9 plots three representations of temperature distribution, using kernel density estimator, histogram and mixture normal distribution respectively. The kernel density estimates are evaluated at 50 points, and the Epanechinikov kernel is used. Kernel density estimates are relatively smooth and are independent of the choice of origin — two desirable qualities that are lacked in the histogram representations. The issue with kernel density estimator is that it is a nonparametric approach so that we cannot parametrically summarize the distribution of

temperature, though it still serves as a visual reference for evaluating other representations.
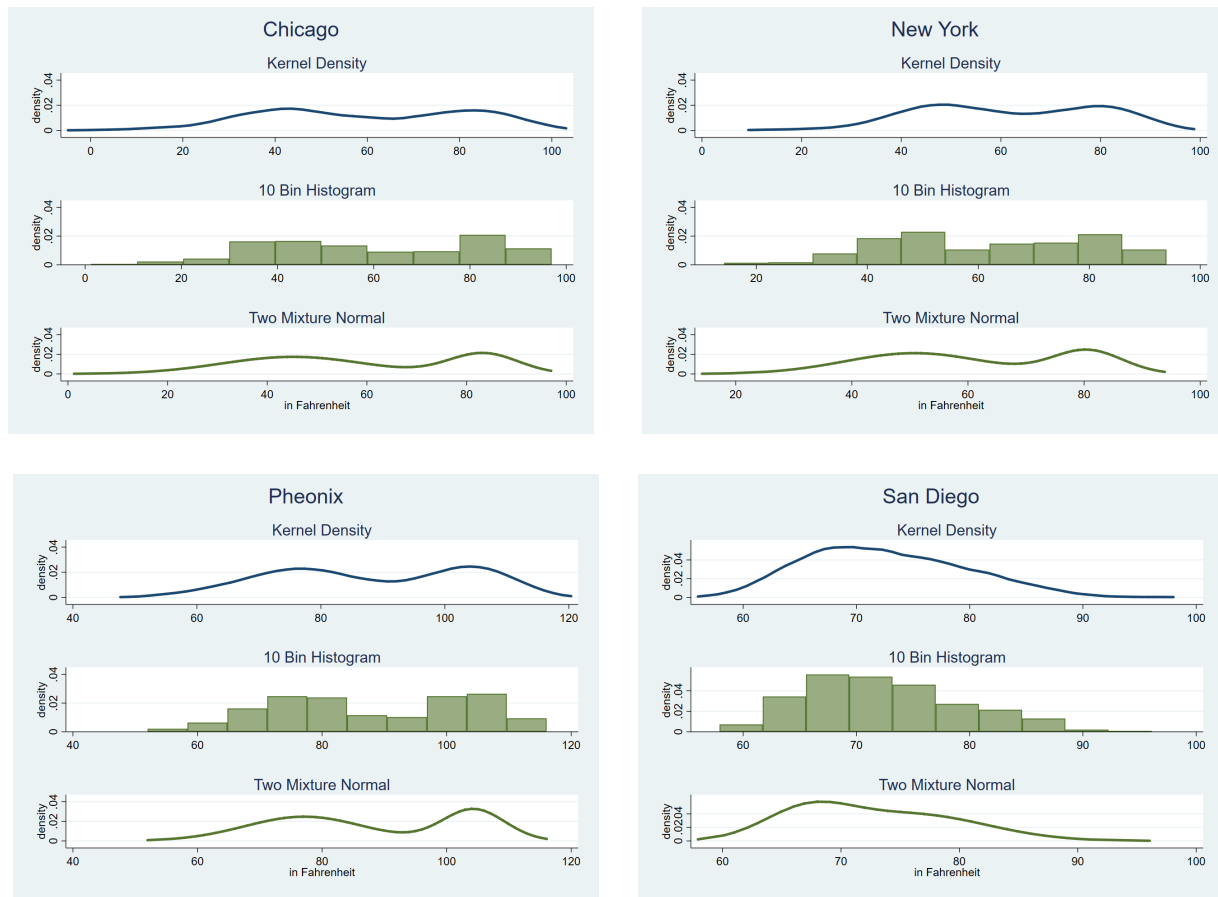

Figure 3.9: Various Representations of Temperature Distribution

Note: These figures plot empirical distributions of daily maximum temperature for Chicago, New York City, Pheonix and San Diego in 2018. The top panels are estimated with kernel density functions; the middle panels are estimated with histograms which have ten bars; the bottom panels are approximated with a two mixture normal distribution

As the kernel density estimates show, the time distribution of temperature follows a bimodal distribution, which indicates that two component mixture normal model could be appropriate in this case. The bottom panels are approximated with two mixture normal distributions that use five parameters: $N_1(\mu_1, \sigma_1^2)$, $N_2(\mu_2, \sigma_2^2)$, and π, the proportion of distribution N_1 (thus the proportion of N_2 is $1 - \pi$). Overall, two mixture normal model is able to capture the key features in the distribution in all of the four examples shown in figure 3.9. There are other more

complicated distribution models for T_{id} that can be used, such as mixture of weibull distribution or other bimodal models.

For the second step, we can estimate $g()$, $h()$ and $k()$ using preferred approaches like Box-Cox, generalized additive models, or neural network model, and select the model that has good prediction power and relatively robust to measurement and prediction error.

The major advantage of this two step approach is that this approach is more adaptive to climate damage projection purpose. Climate change would smoothly shift location and scale parameter estimates, which in turn, would also shift impacts. As noted in section 3.2.4, the histogram representation, which is not a smooth representation, is either insensitive or hyper-sensitive to such shift in weather distributions. Representations of temperature distribution using location-scale distribution could ensure shifts in distribution are well-captured.

Numerical Exercise

We keep the quadratic data generating process as defined in section 3.3.1, summarize temperature distribution with two component mixture normal model, and estimate outcome's response to distribution parameters using both parametric and nonparametric approaches. Figure 3.10 are plots of the real vs. predicted values of outcomes for the testing sample. In this exercise, we focus on three distinct types of models:

1. Ordinary Least Squares. We first model the outcome varable as a linear function of distribution parameters:

$$Y_i = \beta_1 \hat{\mu}_{1,i} + \beta_2 \hat{\sigma}_{1,i} + \beta_3 \hat{\mu}_{2,i} + \beta_4 \hat{\sigma}_{2,i} + \beta_5 \hat{\pi}_i + \varepsilon_i$$

We focus on mean squared prediction error as our key selection criterion. This is constructed

by predicting outcomes using the OLS model fitted with training dataset.

2. Neural Network. Neural networks are non-linear statistical data modelling tools, which can be used to model complex relationships between inputs and outputs. We train a neural network model with two hidden layers to model the nonlinear relationship between outcome and the distribution parameters, and predict outcome with training data.

3. Multivariate Adaptive Regression Splines. The MARS model is a flexible non-parametric regression technique (similar to LASSO), which can be viewed as a piecewise linear model that automatically estimates nonlinearities and interactions among variables. We use the MARS model to explore potential interaction effect between location and scale parameters.

As reported in the first row of table 3.2, the mean squared prediction error for the three models are 235.46, 94.71 and 12.74 respectively, with OLS having the largest mean squared error and MARS the smallest. However, in terms of the overall prediction, a 10 bin regression model has only 6.46 prediction error, which seems to be much better than the alternatives we have proposed in this section.

One interesting observation from figure 3.10 is that as y gets larger, the prediction error for the bin model is more dispersed. This is a natural result following the discussion in section 3.2—problems with the bin model are mostly acumulated at the tail bins. The true DGP is a summation of the daily outcome, which is a quadratic function of daily temperature centered at $10°C$. As a result, a large y value corresponds to having temperature observations being far away from the center. In other words, temperatures are more likely to be widely distributed when y is large. Figure 3.11 is a plot of simulated outcome, and it follows a bimodal distribution. We are interested in predicting for y when y is larger than 80, where temperatures are widely distributed and outcomes are relatively sparse.

Table 3.2 reports mean squared error as we decrease the testing sample size. For example, the row y≥80 corresponds to subseting within the test data for y's that are larger than 80. It is not surprising that the bin model is more preferable in most of these cases, but when y gets to the extremes, MARS model's performance is more stable compared to all other models. Given that the bin regression model has various problems at the tail bins, MARS model seems to be a reasonable alternative that are more adaptive for climate damage projection purpose.
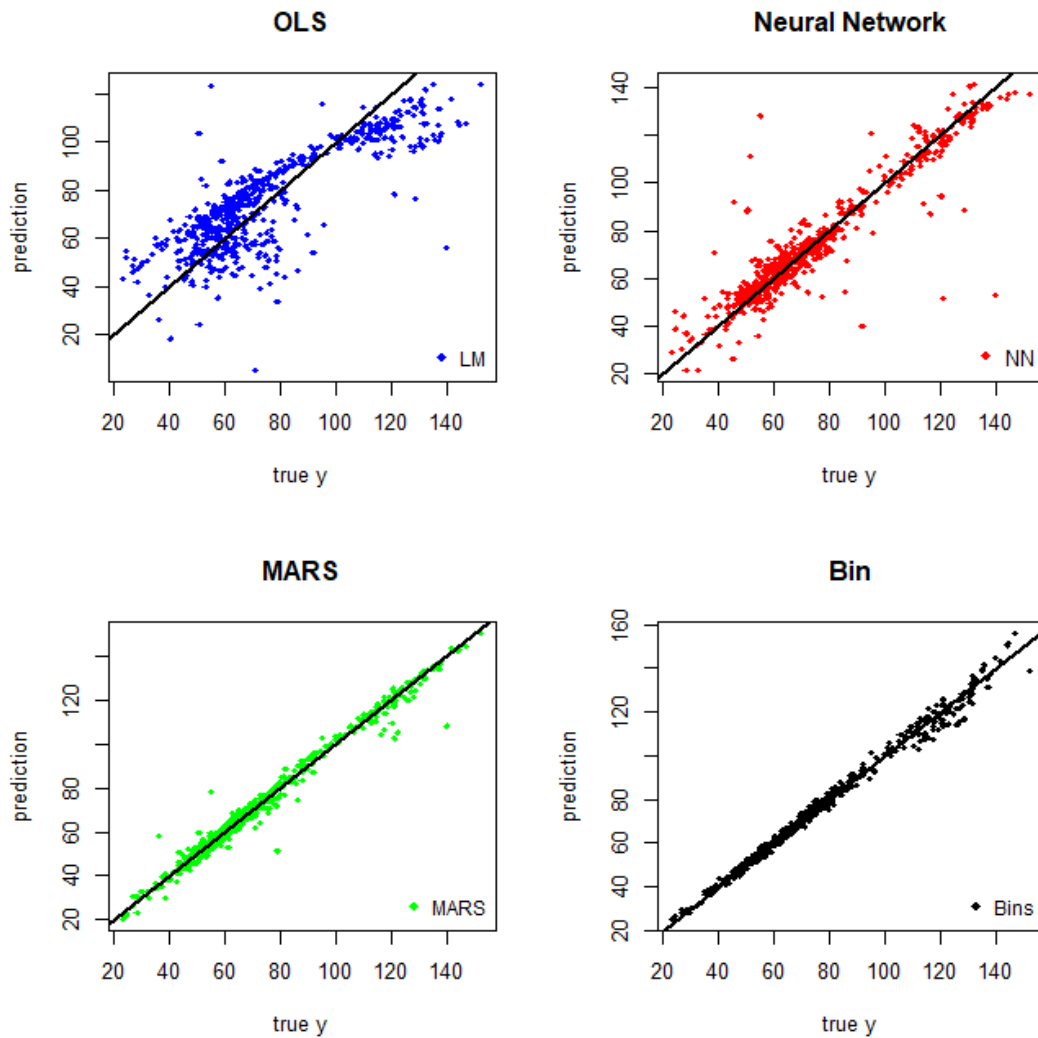

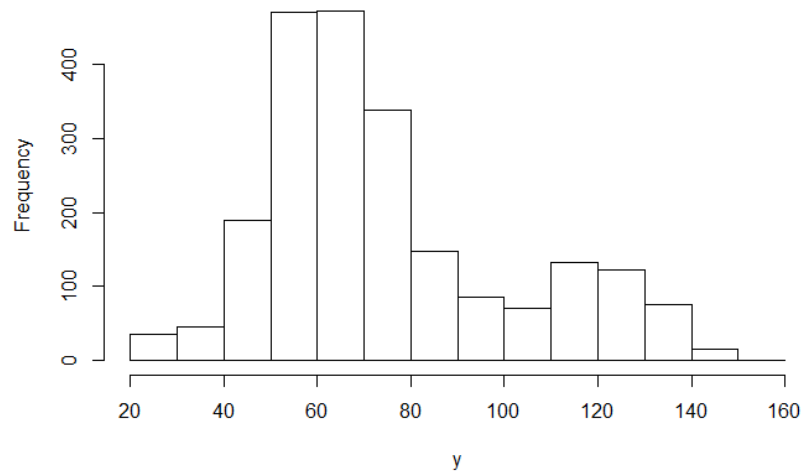**Figure 3.10**: Comparing Prediction Results for Various Models

Figure 3.11: Histogram of Simulated Outcome y

Table 3.2: Mean Squared Error w/ Various Sample Size

	OLS	NN	MARS	BIN
Full sample	235.46	94.71	12.74	**6.46**
$y \geq 80$	331.31	150.39	18.52	**17.04**
$y \geq 90$	377.6	175.85	20.41	**20.28**
$y \geq 100$	404.75	167.98	**22.12**	22.3
$y \geq 110$	481.85	202.21	26.01	**25.79**
$y \geq 120$	699.49	283.84	35.56	**28.76**
$y \geq 130$	985.06	305.48	37.56	**23.54**
$y \geq 140$	1119.09	95.44	**3.9**	72.68
$y \geq 150$	795.84	230.67	**3.7**	188.52

3.4   Conclusion

Binned regression model is heavily used in environmental economics because of its perceived ability to flexibly model impact of changes in the distribution of weather variables. Although in reality, bin model put restrictions on the number of bins, bin width, and slopes within each bin. The estimated bin model is often times not robust to these restrictions. Furthermore,

measurement error interacts with binned model, which leads to misclassification errors that are particularly large at the tail bins. In addition, the binned regression model can be either in- or hyper- sensitive to predicted changes in weather variable, because the nature of binned regression is a histogram representation of density, or a classification approach, which is not smooth.

We explore potential alternatives with better statistical properties. One direction is to pursue more flexible binning, as the methods used in [Schlenker and Roberts, 2009]. We show in the paper that their method are more flexible than exogenously defined binned regression model. In addition, we show that one can also incorporate the weather distribution parameters into regression models, particularly if the main purpose is to look at a smooth shift in the temperature distribution either due to measurement error or for climate damage projection purpose.

3.5   Acknowledgements

Chapter 3 is coauthored with Richard Carson and Dalia Ghanem, and is being prepared for publication. The dissertation author is the principle researcher on this chapter.

3.6   Additional Figures and Tables

3.6.1   Bin Definition: Simulation Design

Let true response $y_{i\tau}$ for location i on day τ be a function of the temperature on that day:

$$y_{i\tau} = f(T_{i\tau})$$

Let $f()$ take the following functional forms:

1. $y = (T - 50)^2 / 1000$

We aggregate $y_{i\tau}$ over years and locations so that $Y_{it} = \sum y_{i\tau}$, $\tau \in t$ and bin the temperature observations. We are interested in the set of β_j when estimating the following regression:

$$Y_{it} = \sum_{j=1}^{J} \beta_j \text{TEMP}_{j,it}$$

where $\text{TEMP}_{j,it}$ represents "number of days in bin j location i in year t". β_j is the bin estimator for $y_{i\tau} = f(\text{T}_{i\tau})$, $T_{i\tau} \in$ bin j. Each β_j would be interpreted as "the effect on Y_{it} of having one additional day in bin j".

The data generating procedure $f(\text{T})$ maps each temperature record T to an outcome variable y, while in bin regression models, the mapping is from a set of temperature observations to an outcome variable y. Therefore, within each bin, we need to choose a point in the bin that can best represent the entire bin, and use that point to compute $f(\text{T})$ for comparison against estimation results. For example, in bin 40°F-50°F, we might pick 45°F as a point that can well represent this bin, and calculate $y = f(45)$ as the true response to a temperature in 40°F-50°F.

There are three candidates for constructing this true response: First, mean temperature of each bin; second, median temperature of each bin; and third, midpoint of each bin. In fact, these three points are almost the same for the middle bins with many observations. Temperature in the middle bins follow almost uniform distributions so that the mean, median and midpoint coincide. However, the mean, median and midpoint at the end bins are far apart since the underlying temperature distributions are skewed. As shown in figure 3.3, the distribution of the lowest bins skew to the right, and the distribution of the highest bins skew to the left. Throughout the paper, we generate true parameter using the median of temperature bins, as median is more robust to skewed distributions. We also conducted simulations using mean and midpoint as alternatives, and they both results in larger misspecification error than using median.

Fix number of observations in each bin. We partition temperature data into 10 bins with equal numbers of observations in each bin. One can see from these figures that middle bins have much smaller bin width while end bins span a wide range of temperatures.

Fix bin widths. We partition temperature data into 10 bins with equal bin width of 10°F. The bin regression model under this bin definition lines up nicely with the true DGP. When bin width is fixed, biases due to misspecification are smaller than fixed observation number, especially at the end bins.

Fix bin widths in the middle and unlimited bounds for end bins. This is the typical setup in the literature where we we partition temperature data into 10 bins; let middle eight bins have bin width of 10°F, and no upper and lower bounds for the tail bins. This bin definition also leads to large misspecification error.

3.6.2   Alternatives to Bin Regression Model: Exponential DGP

We run a similar exercise as in section 3.3 using an exponential function as the true DGP:

$$y = exp(T/50)$$

This exponential function looks quite similar to a linear function in shape. We aggregate the daily outcome variable to station-year level and run the three models as characterized in section 3.3. Figure 3.12, 3.13 and 3.14 are plots of the fitted binned model, Chebyshev polynomial model, and piecewise linear model with bin width equals 1.
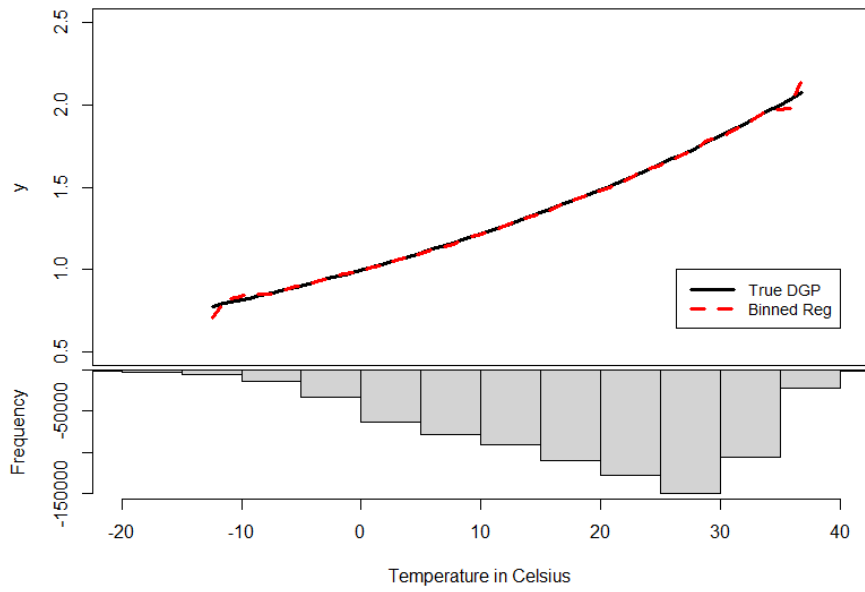
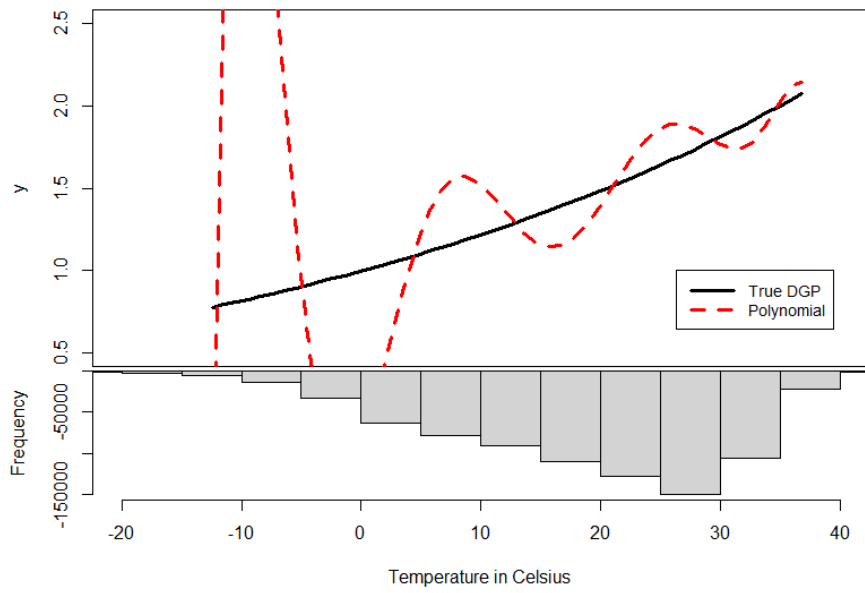**Figure 3.12**: Binned Regression Model (Exponential DGP)

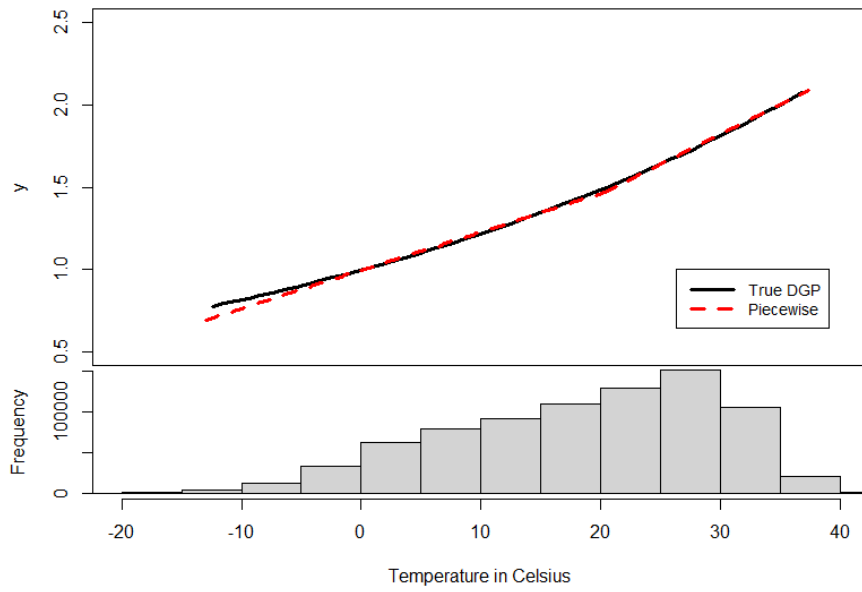**Figure 3.13**: Chebyshev Polynomial Model (Exponential DGP)

Figure 3.14: Piecewise Linear Model (Exponential DGP)

Bibliography

[Ahn and Hamilton, 2019] Ahn, H. J. and Hamilton, J. D. (2019). Measuring labor-force partici-pation and the incidence and duration of unemployment.

[Auffhammer and Aroonruengsawat, 2011] Auffhammer, M. and Aroonruengsawat, A. (2011). Simulating the impacts of climate change, prices and population on california's residential electricity consumption. *Climatic change*, 109(1):191–210.

[Auffhammer et al., 2013] Auffhammer, M., Hsiang, S. M., Schlenker, W., and Sobel, A. (2013). Using weather data and climate model output in economic analyses of climate change. *Review of Environmental Economics and Policy*, 7(2):181–198.

[Barreca et al., 2016] Barreca, A., Clay, K., Deschenes, O., Greenstone, M., and Shapiro, J. S. (2016). Adapting to climate change: The remarkable decline in the us temperature-mortality relationship over the twentieth century. *Journal of Political Economy*, 124(1):105–159.

[Bernhardt et al., 2018] Bernhardt, J., Carleton, A. M., and LaMagna, C. (2018). A comparison of daily temperature-averaging methods: spatial variability and recent change for the conus. *Journal of Climate*, 31(3):979–996.

[Bound et al., 1994] Bound, J., Brown, C., Duncan, G. J., and Rodgers, W. L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 12(3):345–368.

[Bound et al., 2001] Bound, J., Brown, C., and Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics*, volume 5, pages 3705–3843. Elsevier.

[Breusch and Pagan, 1979] Breusch, T. S. and Pagan, A. R. (1979). A simple test for het-eroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294.

[Burke and Emerick, 2016] Burke, M. and Emerick, K. (2016). Adaptation to climate change: Evidence from us agriculture. *American Economic Journal: Economic Policy*, 8(3):106–40.

[Carroll et al., 1995] Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, volume 105. CRC Press.

[Carson et al., 2020] Carson, R. T., Ghanem, D., and Yu, C. (2020). Estimatinig the impact of climate change: An exploration of the bin regression model.

[Carson and Yu, 2020] Carson, R. T. and Yu, C. (2020). Sources and nature of measurement error in estimating climate impacts.

[Considine, 2000] Considine, T. J. (2000). The impacts of weather variations on energy demand and carbon emissions. *Resource and Energy Economics*, 22(4):295–314.

[Dell et al., 2014] Dell, M., Jones, B. F., and Olken, B. A. (2014). What do we learn from the weather? the new climate-economy literature. *Journal of Economic Literature*, 52(3):740–98.

[Deschênes and Greenstone, 2011] Deschênes, O. and Greenstone, M. (2011). Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the us. *American Economic Journal: Applied Economics*, 3(4):152–85.

[Fuller, 1987] Fuller, W. A. (1987). Measurement error models. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*.

[Garg et al., aper] Garg, T., McCord, G. C., and Aleister, M. (working paper). Losing your cool: income and non-income mechanisms in the temperature-violence relationship.

[Graff Zivin and Neidell, 2014] Graff Zivin, J. and Neidell, M. (2014). Temperature and the allocation of time: Implications for climate change. *Journal of Labor Economics*, 32(1):1–26.

[Griliches, 1974] Griliches, Z. (1974). Errors in variables and other unobservables. *Econometrica: Journal of the Econometric Society*, pages 971–998.

[Hausman, 2001] Hausman, J. (2001). Mismeasured variables in econometric analysis: problems from the right and problems from the left. *Journal of Economic perspectives*, 15(4):57–67.

[Hsiang, 2016] Hsiang, S. (2016). Climate econometrics. *Annual Review of Resource Economics*, 8:43–75.

[Hu, 2017] Hu, Y. (2017). The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics. *Journal of econometrics*, 200(2):154–168.

[Lobell, 2013] Lobell, D. B. (2013). The use of satellite data for crop yield gap analysis. *Field Crops Research*, 143:56–64.

[Ma and Guttorp, 2013] Ma, Y. and Guttorp, P. (2013). Estimating daily mean temperature from synoptic climate observations. *International Journal of Climatology*, 33(5):1264–1269.

[Park, 2018] Park, J. (2018). Hot temperature and high stakes exams: evidence from new york city public schools. *URL: https://scholar. harvard. edu/files/jisungpark/files/papernycaejep. pdf*.

[Schennach, 2016] Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, 8:341–377.

[Schlenker and Roberts, 2009] Schlenker, W. and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to us crop yields under climate change. *Proceedings of the National Academy of sciences*, 106(37):15594–15598.

[Scott, 2015] Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

[Shapiro and Francia, 1972] Shapiro, S. S. and Francia, R. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216.

[Wooldridge, 2010] Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.