Mismeasured and Misunderstood: Unmasking the Temperature Proxy Problem in Climate Impact Estimates

Richard T. Carson * Chu (Alex) Yu †

Draft: March 3, 2026

Abstract

Thousands of studies estimate climate change impacts related to temperature, but temperature is rarely measured at the exact location where impacts occur. Empirical work uses proxies such as closest station readings or estimates from gridded data products like PRISM. Measurement error introduced from this practice are routinely ignored. When acknowledged, it is typically assumed small in magnitude and classically (i.i.d. normal) distributed, implying impact estimates are modestly attenuated toward zero. This assumption is formally examined by treating temperature readings at U.S. weather stations as ground truth, then predicting daily mean temperature there using various proxies. This allows precise identification of measurement errors in an otherwise error-free relationship. The results contradict the standard assumption. These errors exhibit strong spatial and temporal components, and, often, are large in magnitude. While proxy measurement errors typically cause downward bias, overestimates are common. Estimation problems are concentrated where attention focuses—the temperature distribution's tails.

Keywords: climate change; measurement error; spatial modeling; temperature proxies
JEL Codes: Q54, C15, C52

*Department of Economics, University of California, San Diego. Email: rcarson@ucsd.edu
†Department of Economics, Wake Forest University. Email: yua@wfu.edu

1 Introduction

People, other biological organisms, and physical processes respond in various ways to changes in climate. The literature documenting these responses has grown exponentially in the last few decades (Grieneisen and Zhang, 2011). Researchers have extensively studied the impacts of climate change on a wide range of outcomes: from mortality (e.g., Barreca et al. (2016)) to productivity (e.g., Graff Zivin et al. (2018)), learning outcomes (e.g., Park et al. (2020)), and mental health (e.g., Mullins and White (2019)). A growing body of research has also explored climate adaptation and acclimatization (e.g., Auffhammer (2022); Carleton et al. (2022); Deschênes and Greenstone (2007); Deschênes and Greenstone (2011); Heutel et al. (2021)). Collectively, these studies have profoundly reshaped our understanding of climate-human interaction and provide useful empirical evidence for relevant climate policies.

An often-remarked-upon but rarely addressed problem in climate economics studies is the deviation between the actual value of the weather variable of interest and the proxy variable (e.g., daily maximum temperature at the nearest monitor) used in the estimated models (Auffhammer et al., 2013; Hsiang, 2016; Stoklosa et al., 2015). This deviation is known as measurement error in weather variables.[1] Such measurement error in climate impact studies is routinely ignored. When acknowledged, it is typically assumed small and classically distributed, leading to the view that a small amount of attenuation bias may be present in climate impact estimates.

In this paper, we challenge this conventional wisdom by formally documenting the nature and sources of errors that are introduced in the typical approaches to constructing temperature proxy variables. We construct a framework in which the validity of the classical measurement error assumption can be empirically tested and find that it is strongly rejected. We focus our attention on how different measurement errors are introduced and the ways in which they influence climate impact estimation. While the analysis centers on temperature, the same structure applies to any stimulus variable measured through spatially and temporally distributed monitors, such as air pollutant concentrations, noise, river flows or wind speed. In each case, researchers must translate point-based measurements into area-level exposures, a process that inherently introduces measurement error.

We start by considering a linear model relating an outcome variable, Y, to true temperature, T^*:

$$Y = \beta_0 + \beta_1 T^* + \varepsilon$$

Unfortunately, T^* is generally not observable. Instead, researchers use a temperature proxy T, which is based on some function, such as the average of readings from a nearby weather monitor, mechanically introducing a measurement error vector u:

[1] We draw on a large literature on measurement errors that explores how they can result in inconsistent estimates of key parameters (e.g., Carroll et al. (2006); Fuller (2009); Griliches (1974); Hausman (2001)).

$$T = T^* + u$$

When estimating the impact parameter β_1 using the constructed proxy T, the OLS estimator $\hat{\beta}_1$ is generally biased. Conventional wisdom holds that measurement errors in such proxies are typically small in magnitude and classical in nature (i.e., distributed i.i.d., with the stronger assumptions of symmetry and normality frequently being made). Under the classical measurement error assumptions, the OLS estimator $\hat{\beta}_1$ converges in probability to the true parameter β_1 multiplied by an attenuation factor, which is a function of the variance of the true temperature and the variance of the measurement error:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 \times \underbrace{\frac{\text{var}(T^*)}{\text{var}(T^*) + \text{var}(u)}}_{\text{Attenuation Factor}}$$

When measurement errors are classical, the attenuation factor is bounded by 0 and 1 so the impact coefficient is biased towards zero, with the magnitude of the downward bias increasing with var(u). If the signal, var(T^*), to noise var(u) ratio is high, this attenuation bias is small. This view of measurement error was explicitly examined by Lobell (2013), who conducted a simulation exercise on measurement error in temperature and precipitation variables. He finds adding a small amount of classical measurement error to the temperature variable only slightly affects the coefficient measuring crop yield response to temperature. This is the expected result of adding a small amount of classical measurement error.

However, the conditions underlying the classical measurement error assumption are highly restrictive and have not been comprehensively examined in climate impact literature. If those measurement errors are not classical, then the direction of the bias is uncertain. If they are not small, then the magnitude of the bias becomes uncertain.

We formally document the nature and sources of errors introduced in the typical approaches to constructing temperature proxy variables. Measurement error in a temperature variable proxy can be decomposed into different sources. Section 2 lays out how this occurs for monitor interpolation, temporal interpolation, spatial aggregation, temporal aggregation, and measurement error's interaction with model specification.

To empirically examine the properties of temperature measurement error, we conduct a novel quantitative exercise in Section 3 using high-quality data from a large network of U.S. weather stations. We designate the observed daily mean temperature T_{mean} at each weather station as the ground truth, simulating the ideal scenario of having a temperature monitor at each location of interest.[2] Our approach establishes a simple, deterministic data generating

[2] A point-level ground truth exists only for some error sources. Accordingly, Section 3 focuses on monitor interpolation, temporal interpolation, and binning, while aggregation-related errors are examined empirically in Section 4, where no meaningful area-level "true" temperature exists.

process where the outcome variable Y equals this true temperature: $Y = 1 \times T_{mean}$.[3] We then systematically construct various temperature proxies that incorporate different sources of measurement error. The key step in our exercise is regressing the actual temperature T_{mean} on these constructed proxies. In a scenario with no measurement error, this regression would yield a slope coefficient of exactly 1 and an R-squared of 1. This setup provides a natural interpretation: when we regress the true temperature at a location on its proxy measure, the deviation of the estimated coefficient from its true value of 1 directly reveals the direction and magnitude of bias, while the R-squared values summarize how much true temperature variation is captured by each proxy.

Different proxies introduce distinct measurement error components. We examine commonly used types of temperature proxies. The first type is based on formulaic calculations using readings from one or more nearby monitors. This approach commonly uses inverse distance weighting, where monitors closer to the location of interest receive greater weight. For simplicity, we focus on the most basic version of this approach - using readings from only the nearest monitor as the proxy, which effectively assigns a weight of 1 to the closest monitor and 0 to all others. The second type involves adaptive smoothing (e.g., kriging) of monitor readings, incorporating topographical features such as elevation and distance from large water bodies. These are often referred to as "gridded data products". We use PRISM as our example of this approach, as this dataset has been widely employed in economic studies of climate impacts in the U.S (e.g., Baylis (2020); Deschênes and Greenstone (2007)). For these proxies, we show that measurement errors are often large in magnitude and exhibit systematic patterns across space and time, violating the classical i.i.d. assumption. For instance, we find that measurement errors tend to be larger in regions with complex terrain, such as the Rocky Mountains, and during winter months when temperature variability is higher.

In addition to the numeric exercise characterizing measurement error across temperature proxies at the station level, we examine how these errors translate into aggregation bias in an applied setting by revisiting the well-documented relationship between temperature and crime from 2010–2019 (e.g., Berman et al. (2020); Heilmann et al. (2021)). We systematically evaluate the consequences of spatial aggregation, temporal aggregation, and their interaction with the binned regression framework. To foreshadow results, we find biases for different aggregation schemes ranging from -30% to 300%.

Our analysis yields several important insights for empirical researchers working with temperature data.[4] First, we demonstrate that sophisticated gridded data products like PRISM do not generally outperform simpler approaches—when monitoring networks are sufficiently dense, as in many urban settings, nearest-neighbor interpolation often produces less measurement error. Second, building on earlier results that panel fixed effects can reduce signal-to-noise ratios (Dong

[3]The coefficient on T_{mean} is set "1" to facilitate easy interpretation. Any known non-zero constant works.

[4]While temperature is the weather variable most often used in climate impact estimation, others such as precipitation involve analogous measurement issues.

and Millimet, 2024; Fisher et al., 2012), we show that widely-used, fixed-effect specifications can substantially amplify the adverse influence of temperature measurement error. Third, we find that measurement errors are particularly severe at the extremes of the temperature distribution, with high temperatures typically underestimated and low temperatures typically overestimated. This pattern warrants special caution when employing nonlinear models, as parameters in such specifications can interact with these non-classical measurement errors in ways that exacerbate bias in precisely the temperature ranges of greatest interest.

Our paper makes two main contributions to the literature. First, we formally document and characterize the nature of temperature measurement error in climate impact studies. While measurement error issues have been well documented in the broader econometrics and labor economics literature (e.g., Black et al. (2000); Fuller (2009); Griliches (1974)), and have been examined in other environmental contexts where the underlying stimulus is measured with error, such as air pollution data (Fowlie et al., 2019), and remotely sensed data (Alix-Garcia and Millimet, 2023; Josephson et al., 2024; Proctor et al., 2023), measurement errors in temperature and similarly precipitation variables have not been systematically documented. While measurement error in climate variables has been acknowledged (e.g., Auffhammer et al. (2013); Hsiang (2016)), the literature lacks a comprehensive analysis that documents the full range of temperature measurement errors arising from commonly used proxies. Some climate studies have developed solutions for particular challenges - such as weighting schemes for complex spatial aggregation (e.g., Burke et al. (2018); Carleton et al. (2022); Schlenker and Roberts (2009)). Our analysis complements this work by documenting these errors in a unified way and giving practical guidance in Section 5.

Second, we empirically quantify the magnitude, distribution, and impact of measurement errors using a novel approach that treats weather station data as ground truth and constructs a deterministic relationship with known parameters. Our empirical strategy allows us to isolate and exactly identify measurement errors in an otherwise errorless relationship, providing a rigorous assessment of how different proxy construction methods affect climate impact estimates. The analysis examines how these error components interact with model specifications commonly used in climate economics, including fixed effects and nonlinear functional forms. While previous studies have noted that panel fixed effects might reduce signal-to-noise ratios (Dong and Millimet, 2024; Fisher et al., 2012), our analysis provides comprehensive quantification of this interaction.

The paper is organized as follows. Section 2 provides background on measurement error models and systematically outlines the underlying sources of temperature measurement error. Section 3 presents the framework of the quantitative exercise used to isolate proxy measurement error. We then use this framework to empirically characterize the properties of the identified measurement error and formally test whether it satisfies the conditions underlying the classical measurement error assumption. Section 4 examines how measurement errors affect estimates of the climate-crime relationship. Section 5 provides practical guidance for empirical researchers

and outlines prospective uses of our framework.

2 Background

In this section, we lay out the theoretical framework for analyzing temperature measurement error. We first review the classical measurement error model and its key assumptions. Then we identify and categorize five distinct sources of measurement error that arise when constructing temperature proxies. Finally, we briefly review related literature on measurement error in economics and climate studies to help place our analysis in context.

2.1 Measurement Error Models

We begin with the most basic linear model for estimating the influence of temperature on an outcome variable of interest. In that model, daily mean temperature is the only independent variable, and interest lies in estimating the coefficient β_1 in the outcome variable y_{id}:

$$y_{id} = \beta_0 + \beta_1 T^*_{id} + \varepsilon_{id} \tag{1}$$

where T^*_{id} represents the true daily mean temperature at location i and time d. Under ideal circumstances, this model would be estimated using high frequency, evenly-spaced temperature readings from high-quality monitors located exactly at each agent's individual location i.[5] No measurement error is assumed to present in this situation.[6] However, in practice, constructing the temperature exposure at a given location and point in time inevitably introduces a measurement error component, u_{id}. Assuming this error is additive, the observed (proxy) temperature T_{id} can be expressed as $T_{id} = T^*_{id} + u_{id}$.

In the classical measurement error model, the following conditions are assumed to hold:

$$\begin{aligned}
&\text{A1.} \quad u_{id} \sim i.i.d.(0, \sigma_u^2), \ \forall i, d \\
&\text{A2.} \quad \text{cov}(T^*_{id}, u_{id}) = 0 \\
&\text{A3.} \quad \text{cov}(\varepsilon_{id}, u_{id}) = 0 \\
&\text{A4.} \quad E(u_{id}^2) < \infty
\end{aligned}$$

We formally test these classical assumptions in Section 3.3 through a detailed analysis of U.S. weather station data.

To foreshadow our results, we demonstrate that empirically constructed measurement errors deviate substantially from assumption A1: they exhibit non-zero means, strong serial

[5] High frequency, evenly-spaced temperature readings at the monitor allow for mean daily temperature to be calculated without appreciable error.

[6] We are implicitly assuming that monitoring equipment functions perfectly. In the real world applications, temperature datasets often include quality flags to identify records impacted by equipment malfunction or other known measurement problems. We follow the standard practice of dropping such flagged observations.

correlation, heteroskedastic variance, and in some cases mild asymmetry. Assumption A2 posits that true temperature and measurement error are uncorrelated. Instead, measurement errors show strong correlation with the magnitude of the true temperature. Since the estimation procedure matches the true data generating process, assumption A3 is not testable in our setting.[7] A4 is the usual (weak) finite variance requirement and is never empirically violated. Some definitions of classical measurement error (Carroll et al., 2006; Fuller, 2009) included the stronger requirement of symmetry or, more typically, normality. We reject these assumptions as well.

By substituting the expression for the constructed proxy variable $T_{id} = T^*_{id} + u_{id}$ into equation (1), we obtain:

$$\begin{aligned} y_{id} &= \beta_0 + \beta_1(T_{id} - u_{id}) + \varepsilon_{id} \\ &= \beta_0 + \beta_1 T_{id} + e_{id} \end{aligned}$$

The new error term $e_{id} \equiv \varepsilon_{id} - \beta_1 \times u_{id}$ introduces bias due to endogeneity in the estimation of β_1. In general, when assumptions A3 and A4 are satisfied, the effect of measurement error on estimation can be written as:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 \times \frac{\text{var}(T^*) + \text{cov}(T^*, u)}{\text{var}(T^*) + \text{var}(u) + 2\,\text{cov}(T^*, u)} \tag{2}$$

From equation (2), when assumption A2 is not met, measurement errors are correlated with true temperatures. Because the sign of $\text{cov}(T^*, u)$ is not restricted, the direction of bias becomes ambiguous. A priori, it is not hard to conjecture that this covariance term could have different signs. As we will show in Section 3.3, this correlation can be substantial and it can lead to overestimation rather than attenuation when $\text{cov}(T^*, u)$ is negative and large enough.

When all classical measurement error assumptions hold, including A2: $\text{cov}(T^*, u) = 0$, the OLS estimator simplifies to:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 \times \frac{\text{var}(T^*)}{\text{var}(T^*) + \text{var}(u)} \tag{3}$$

The quantity $\frac{\text{var}(T^*)}{\text{var}(T^*)+\text{var}(u)}$ is the ratio of the variance of signal to total (signal plus noise) variance, which is sometimes referred to as the reliability ratio. The value of this term, by construction, always falls between 0 and 1. Thus increasing the level of noise always increases the attenuation of the estimated coefficient towards zero. When the variation of measurement error is relatively small compared to the variation of true temperature ($\text{var}(u) << \text{var}(T^*)$), the term $\frac{\text{var}(T^*)}{\text{var}(T^*)+\text{var}(u)}$ is dominated by $\text{var}(T^*)$, and thus the attenuation bias due to measurement error is minimal.

[7] Assumption A3 requires measurement error to be uncorrelated with the error term in equation (1), which relates to correct model specification. Our analysis simplifies this by using a known data generating process with perfect specification matching. In practice, A3 violations typically increase measurement error magnitude and are context-specific (e.g., sparse air pollution monitoring in areas with limited healthcare access), but are unlikely to counterbalance A1 and A2 violations to produce classical measurement error. We do not explore A3 violations.

Conversely, when the variation of measurement error is large compared to the variation of true temperature ($\text{var}(u) >> \text{var}(T^*)$), β_1 shrinks toward zero.

In equation (3), the temperature proxy represents the only source of signal. However, this changes when spatial and/or temporal fixed effects are included in the model. The interaction between fixed effects and temperature proxies can be understood through the implementation of fixed effect models: one approach involves regressing the dependent variable on the fixed effects, extracting the residuals, and then regressing these residuals on the remaining covariates. When fixed effects capture variation that helps predict true temperature, the usable variance in the temperature proxy signal is mechanically reduced. This reduction amplifies attenuation bias under classical measurement error assumptions. In Section 3.4, we provide empirical estimates that quantify the fraction of daily temperature variance absorbed by different sets of location- and time-based fixed effects.

2.2 Types of Measurement Errors

Construction of temperature proxy variables typically involves multiple steps, each introducing a different type of measurement error. Here, we elaborate on the potential sources of these errors, denoted as u_{id} in the construction of temperature proxies $T_{id} = T^*_{id} + u_{id}$. Two fundamental types of measurement error - monitor interpolation and temporal interpolation - are always present when constructing temperature proxies. Unlike other sources we discuss below, these two can be directly constructed and tested against classical measurement error assumptions since we have access to ground truth temperature data at weather stations.

2.2.1 Monitor interpolation error

Monitor interpolation error occurs because weather monitors rarely coincide with locations of interest (e.g., individuals or ecosystems). Instead, temperature proxies are constructed using readings from one or more nearby monitors. With advancements in computational power, researchers often average readings from multiple monitors, either based on proximity or within defined radii. Techniques range from using the nearest monitor reading to more sophisticated methods like kriging that use readings from multiple monitors as inputs. While these advanced methods can sometimes reduce interpolation error, they cannot eliminate it entirely and, as we demonstrate, may exacerbate it.

2.2.2 Temporal interpolation error

Temporal interpolation error arises when the outcome of interest (e.g., the number of daily recreational trips to beaches, golf courses, or ski resorts) is influenced by continuous temperature variation, but the proxy (e.g., daily/weekly minimum or maximum temperatures) fails to capture

this variation adequately. If we assume a continuous and linear relationship between temperature and behavior, then the daily mean temperature (T_{mean}) serves as a sufficient summary statistic.

However, the way T_{mean} is calculated can lead to significant discrepancies. A common approach involves averaging the daily maximum (T_{max}) and minimum (T_{min}) temperatures. This method's popularity stems from historical reporting standards where monitors typically recorded only these two values (e.g., GHCN Daily[8]). Studies using long time periods are effectively constrained to this particular T_{mean} proxy, and it has been incorporated into widely-used temperature datasets such as PRISM.

While this averaging method appears intuitive, it introduces several problems. First, it provides an unbiased estimate of T_{mean} only if temperature follows a symmetric distribution throughout the day—an assumption that, as we demonstrate, typically does not hold in practice. Second, even when the symmetry condition is met stochastically, averaging these two extreme order statistics to estimate the mean produces one of the noisiest possible unbiased estimates. Third, the relationship between T_{min}, T_{max}, and T_{mean} varies seasonally. From a physical perspective, the Earth's orbital position influences the timing of daily temperature extremes, generating systematic seasonal patterns in measurement errors characterized by strong temporal autocorrelation.

Alternatively, T_{mean} can be calculated as the hourly average of 24 temperature readings, which provides a more accurate representation. This method, considered the "true" T_{mean}, diverges minimally from minute-by-minute measurements (Ma and Guttorp, 2013). Our empirical tests reveal that using the common approach of averaging daily maximum and minimum temperatures introduces an additional error component compared to monitor interpolation alone.

2.2.3 Spatial aggregation

Spatial aggregation errors occur when outcomes are measured over regions exposed to varying temperatures. For example, daily violent crime counts in U.S. counties are often aggregated over areas with different temperatures. Spatial aggregation errors become more significant as the area size increases. In a county like San Diego, California, for instance, the daily maximum temperature on the coast may be 40°F lower than at its eastern desert edge 86 miles away. In contrast, temperature variation across Manhattan (New York County) is minimal.

These observations have two key implications. First, using a single temperature measure to represent a large region can introduce significant measurement error. Second, the magnitude of this error varies in predictable ways with location, challenging the assumption of independent and identically distributed (i.i.d.) errors (Assumption A1). This error is also likely correlated with other aspects of the model, such as the precision of the dependent variable, which itself varies with the number of units aggregated (Assumption A3).

[8]The Global Historical Climatology Network daily (GHCNd) is an integrated database of daily climate summaries from land surface stations across the globe.

To further illustrate spatial aggregation, consider a county with a uniform distribution of agents (e.g., people or farms), all potentially exposed to different temperature conditions. In the model, temperature is recorded at the county's geographic centroid. If temperature follows a symmetric gradient and responsiveness is linear, errors may cancel out, resulting in minimal bias in the temperature slope estimates across counties. However, when responsiveness is quadratic, these errors no longer cancel, and bias increases with greater temperature variation across the region. Additionally, factors correlated with county size, such as elevation or distance from physical features like coasts or forests, can further complicate temperature measurement, making a smooth, well-defined gradient less likely.

The challenge becomes most pronounced in larger counties, where the uniformly distributed agent assumption becomes untenable. For instance, in major coastal counties in the western U.S., much of the population is concentrated near the coast, meaning the population centroid is often much further west than the geographic centroid. Conversely, agricultural production might be concentrated inland, creating the opposite disparity. These examples highlight the difficulties of assuming that measurement error is uniformly distributed and uncorrelated with other aspects of the model, such as county size or the precision of dependent variables. Some of these challenges have been substantially mitigated in recent empirical studies through improved spatial weighting and aggregation practices. This issue is discussed in more detail in Section 5.1.

2.2.4 Temporal aggregation

Temporal aggregation errors arise when a single temperature is used to represent an extended time period. For example, in studies of monthly electricity bills, researchers often rely on average monthly temperatures, which overlook daily fluctuations. A month with a steady temperature of 65°F is treated the same as a month with alternating daily mean temperatures of 40°F and 90°F, despite very large differences in energy usage among U.S. households.

To address this issue, researchers have developed metrics like cooling degree days, which sum the degrees above a certain threshold (e.g., 65°F). However, these measures assume a linear relationship between temperature and response, which is often not realistic. For instance, households are less likely to adjust their thermostats when the temperature shifts from 70°F to 75°F compared to a shift from 80°F to 85°F, indicating a nonlinear response. Agricultural production models, such as those studied by Schlenker and Roberts (2009), illustrate further complications. A farm's output may be influenced differently by temperature during planting, cultivation, and harvesting stages. Similarly, temporal aggregation error can arise when a daily mean temperature is used to explain outcomes influenced by specific temperature moments, such as an athletic event or a crime occurring at a particular time of day. In time series analyses, the problem is compounded by a lack of understanding of how temperature affects outcomes over different intervals. At one extreme, only temperature in a narrow time window may matter; at

the other, detailed temperature data over quite long periods is required (Kolstad and Moore, 2020).

Most empirical studies rely on panel datasets with repeated observations over time across various locations (Ghanem and Smith, 2021). While panel data offers many advantages, it also presents challenges related to spatial and temporal aggregation, especially when outcomes are aggregated over large regions or extended time intervals. For instance, outcomes like county-level infant mortality rates may suffer from spatial aggregation issues, while temporal aggregation issues arise when the outcome is measured on a coarser time scale than the temperature stimulus.

Finally, relaxing the assumption of a linear response introduces additional complexity. Many climate studies assume nonlinear relationships, particularly in the tails of the distribution. Models like binned regressions, kernel density regressions, and splines allow for flexible responses to temperature extremes. However, this flexibility can amplify the influence of measurement errors in the proxy. To illustrate this issue, we focus on one popular approach, the binned regression model (e.g., Barreca and Schaller (2020); Burke et al. (2018); Deschênes and Greenstone (2011)), which can be seen as a piecewise linear regression on a series of bin indicator variables created for different temperature ranges. Bins are designed to enhance robustness by discarding some temperature information, treating all observations within the same bin as exposed to identical temperature stimuli. However, measurement error may cause observations to be misclassified into incorrect bins. We demonstrate that while this misclassification is symmetrically distributed in central bins, it becomes increasingly one-sided and pronounced as one moves toward the tail bins. While the interaction between the binned regression model and measurement error is particularly easy to illustrate, other nonlinear models that gain flexibility by adding parameters through some form of series expansion—and that effectively place greater weight on tail observations—are likely to exhibit qualitatively similar behavior.

2.2.5 Model-based spatial smoothing products

The PRISM dataset has been extensively used in studies estimating temperature and climate impacts in the US. It provides daily gridded maximum and minimum temperature data for the continental United States at a 4km grid cell resolution.[9] PRISM offers a sophisticated solution to the monitor interpolation problem.[10] It calculates a climate-elevation regression for each grid cell, assigning weights to stations based on their physiographic similarity to the grid cell (Daly et al., 2008). This method considers factors such as distance to the grid cell, elevation, coastal proximity, topographic facet orientation, vertical atmospheric layers, topographic position, and

[9]Earlier studies using PRISM include Deschênes and Greenstone (2007), which take month-by-year PRISM measure of precipitation and temperature to study the effect of temperature and precipitation on agricultural profits; Schlenker and Roberts (2009) downscaled PRISM monthly temperature data to daily level heat exposure levels, with this downscaled dataset also being widely used in subsequent research, such as by Colmer and Doleac (2023); Hendricks (2018); Mullins and White (2019).

[10]PRISM's measurement error is not conceptually distinct from monitor interpolation. We present it separately for expositional clarity.

the orographic effectiveness of the terrain.[11] Despite the advanced algorithm, nothing forces PRISM dataset's measurement errors to be independent and identically distributed. Indeed, the substantial differences in the information underlying PRISM temperature estimates across locations help explain why the A1 measurement error assumption is often violated.

To align the PRISM dataset with outcome variables of interest, different interpolation and aggregation schemes are still necessary. For example, in Klaiber et al. (2017)'s study on the tradeoff between temperature and green amenities, they averaged over July's minimum temperature across all housing transactions in each tract/year combination to create their census tract level temperature measures. Baylis (2020) aggregated the gridded data to the CBSA level using population weights to examine the relationship between temperature and expressed sentiment on Twitter; Auffhammer (2022) matched the PRISM grids to ZIP code shapes from the census. Heutel et al. (2021) used NOAA's GHCN weather data for their main analysis and PRISM as a robustness check, finding qualitatively comparable results between the two, although the impact magnitudes are much smaller using PRISM, which they attributed to PRISM's tendency to assign fewer extreme values.

2.3 Related Literature

Measurement error is a core problem in empirical economics. Its presence leads to inconsistent parameter estimation, which can result in incorrect policy implications to varying degrees. Not surprisingly the implications of measurement error have been extensively studied in econometrics and statistics literature. Key works include Carroll et al. (2006); Fuller (2009); Griliches (1974) and Hausman (2001) among others. Recent theoretical developments in estimation methods for nonlinear models in the presence of measurement errors are reviewed by Chen et al. (2011) and Schennach (2016).

Many issues described in Section 2.2 also arise elsewhere and are documented in the broader econometrics literature, although often with a different, and typically, more general, nomenclature. For instance, the monitor interpolation issue is referred to as "spatial misalignment" in the econometrics literature, to denote measurement error problems that can arise when geographic boundaries defining the independent and dependent variables do not coincide. In a recent paper, Pouliot (2023) proposes two new complementary methods to account for this misalignment. They conduct a detailed reanalysis of Maccini and Yang (2009)'s early-life rainfall paper and finds economically significant quantitative differences, but the more qualitative conclusions largely hold up.

Temporal aggregation is sometimes referred to as "mixed data sampling" (MIDAS), when the unit of time between covariates and dependent variable do not align (Andreou et al., 2010;

[11] Conceptually, if PRISM used only distance of monitors from the grid cell centroid and had deterministic weights that declined in a linear fashion with that distance, then it would be an inverse distance weighting interpolation estimate. It differs from that by empirically estimating those weights using a smoothing approach and using various physical science factors known to influence spatial temperature patterns.

Khalaf et al., 2021; Yang et al., 2023). Andreou et al. (2010) raises the point that there is no a priori reason why one should (i) ignore the fact that the variables involved in empirical models are in fact generated from processes of different/mixed frequencies and (ii) estimate econometric models based on an aggregation scheme of equal weights. Spatial aggregation is another version of MIDAS where the geographies are at different "frequencies" instead of the time units being at different frequencies.

In the field of labor economics, creative use of multiple proxies has often been found to provide a reasonable solution to measurement error problems (Bernal et al., 2016; Black et al., 2000; Black and Smith, 2006; Lubotsky and Wittenberg, 2006). For example, Black and Smith (2006) use multiple proxies of college quality and various econometrics approaches to correct for measurement error in college quality, showing existing literature understates the wage effects of college quality. Lubotsky and Wittenberg (2006) examine the relationship between parents' income and children's education outcomes and demonstrate that the reduction in attenuation bias from a better use of proxy variables can be significant. The multiple proxy approach has also seen recent use in the examination of data from economic experiments (Gillen et al., 2019).

From the general econometrics and statistics literatures on measurement error, it might be tempting to think an instrumental variable approach would be a good solution to the measurement error problem we have described involving temperature proxies. Specifically, since there are multiple ways of measuring/predicting temperature at a given location/time and one can use one reading as the instrument for the other reading. Empirical evidence presented in Section 3 strongly suggests that the exclusion restriction necessary for instrument validity does not hold. Specifically, measurement error in temperature variables follows strong seasonal and spatial patterns that are highly correlated with the true temperature reading. Additional discussion of the IV approach is contained in Appendix C.1.

Measurement error in environmental measures has been examined in a variety of related contexts. For example, Fowlie et al. (2019) document spatial mismeasurement in regulatory air pollution monitors, showing how limited monitoring coverage can distort exposure estimates. Proctor et al. (2023) demonstrate that multiple imputation methods can reduce bias and improve coverage when using such remotely sensed data and Alix-Garcia and Millimet (2023) provide a framework for correcting misclassification in binary satellite-derived outcomes. Josephson et al. (2024) further shows that different earth observation products measuring the same weather conditions produce systematically different estimates, with economic findings sensitive to the data source.

3 Assessing Temperature Measurement Error Properties

We start by asking what an ideal experiment would look like to assess how measurement error in a temperature proxy affects the estimated relationship of interest. Our approach parallels

early econometric exercises that evaluate parameter recovery under known DGP and related macroeconomic calibration studies that test whether models reproduce key empirical moments (e.g., Bound et al. (1995); Kydland and Prescott (1982); Staiger and Stock (1997)). The key element in these efforts is the elimination of extraneous factors that can hinder the assessment of a specific relationship.

The usual challenge in estimating climate-related effects is that researchers rarely know the true form of the response to temperature—whether it is linear, nonlinear, discontinuous, heterogeneous across agents, or driven by cumulative exposure. We abstract from these complications by constructing an idealized natural experiment in which "truth" is defined as temperature readings from a ground monitor at a fixed latitude–longitude point. This design allows us to characterize proxy-level measurement error in isolation, independent of the spatial or temporal aggregation choices that arise when fine-scale temperatures are mapped to coarser units (an issue addressed separately in Section 4).[12]

3.1 Measurement Error Construction

3.1.1 An Idealized Natural Experiment

Our quantitative exercise starts by treating readings from high-quality U.S. weather stations at specific locations as the ground truth that we seek to predict. We then systematically look at the quality of predictions of this ground truth obtained by using temperature proxies constructed in different ways. By design, our response variable contains no true error component. This ensures that any deviation of the slope estimate from its true value "1" or the deviation of R-squared value from "1" stems solely from measurement error properties induced by the particular temperature proxy construction method.[13] In our numerical exercise, the R-squared statistic directly represents the proportion of true temperature variation explained by the proxy measure. The deviation of estimated β_1 from its true value of "1" measures the bias in the impact estimates induced by the proxy.

Having identified the measurement errors, we test whether they conform to classical measurement error assumptions. The panel structure of our data allows us to examine spatial and temporal patterns in measurement errors. Such patterns represent particularly concerning deviations from classical measurement error assumptions because they indicate potential correlation with other factors of interest in empirical climate research.

[12]This feature provides a rare advantage relative to other literature, such as income and wealth measurement, where even benchmark administrative data are subject to non-classical measurement error (Millimet and Parmeter, 2025).

[13]In this setup, the regression is simply using a proxy temperature measure to predict the true monitor reading at that same location. Under perfect measurement—i.e., when the proxy equals the truth—the regression of T^* on the proxy yields a slope of exactly 1 and an R-squared of 1. We normalize the true slope to 1 for expositional convenience; any non-zero constant would produce the same logic. Many climate impact studies either assume linear temperature responsiveness or embed linearity as a special case of more flexible nonlinear specifications, so this normalization is without loss of generality.

Assumed Data Generating Process. We perform the quantitative exercise separately for each weather station in our sample.[14] For each monitor, we specify the following DGP:

$$y_d = 1 \times T_{mean,d}$$

where $T_{mean,d}$ represents the arithmetic mean of 24 equally spaced hourly temperature observations on day d at that monitor. We treat $T_{mean,d}$ as the ground truth temperature. By deliberately excluding a stochastic error term, we ensure that any deviations are solely attributable to measurement error.

To assess the influence of measurement error, we use OLS to estimate the following:[15,16]

$$y_d = \beta_0 + \beta_1 T_d + \epsilon_d \tag{4}$$

where T_d is the proxy temperature measure. Since we have constructed the DGP as $y_d = T_{mean,d}$, estimation of equation (4) has a natural interpretation: predicting true temperature using a proxy measure for it. Without measurement error (when $T_d = T_{mean,d}$), this regression yields an intercept of zero, a slope coefficient of 1, and an R-squared of 1.

Key Statistics for Analysis. Two sets of statistics are of primary interest. The first set are summary measures from equation (4), such as the loss in predictive power from having to use the proxy and the relative bias in the slope coefficient. The second set are test statistics related to whether the constructed measurement error is consistent with classical measurement error assumptions. The panel structure of our data – temperature observations across multiple locations and time periods – allows us to examine whether there are spatial and temporal deviations from the i.i.d. assumptions that characterizes classical measurement error.

Temperature Proxy Construction. We consider two main approaches to proxy construction, starting with the simplest and most commonly used method: readings from the nearest available monitoring station. When combined with calculating daily mean temperature as the average of daily minimum and maximum temperatures, we refer to this as "traditional interpolation". At the opposite end of complexity, we examine PRISM, a sophisticated smoothing model incorporating topographical features. While intermediate methods such as inverse dis-

[14] In Section 3.4, we complement this station-level analysis with an aggregated analysis that pools observations across all weather stations to examine broader patterns of measurement error.

[15] Note that our measurement error framework in the introduction $T = T^* + u$ defines the structure of measurement error throughout the paper. Equation (4) serves a different purpose - quantifying the impacts of measurement error - rather than redefining the measurement error itself.

[16] Cramer (1987) shows that the sample R-squared is upwardly biased relative to its population value in finite samples. Our sample sizes are sufficiently large that this effect should be small.

tance weighting (IDW)[17] with multiple monitors exist between these extremes, our analysis focuses on these two endpoints, as they represent the range of approaches commonly used in the literature.

Use of Fixed Effects as an Alternative Proxy Specification. Our experiment setup uses a panel of monitoring stations across locations and time periods, allowing us to test measurement error's correlation with spatial (county, state) and temporal variables (month, day). These spatial and temporal variables can serve as alternatives to the temperature proxies previously discussed. For example, location-specific fixed effects would capture the mean temperature at each monitor location; location-by-month fixed effects would more accurately recover the mean temperature at each location in each month. At the far extreme are the climate normals – day-of-year indicators for each location that effectively capture daily temperature patterns throughout the year at specific locations.[18]

The typical use for such fixed effects in empirical climate studies serves a different purpose: they primarily control for unobserved heterogeneity by purging correlations between idiosyncratic aspects of individual units (e.g., households or farms) or temporal patterns of economic activity and temperature. While fixed effects help control for confounding factors, they also simultaneously absorb substantial variation in the temperature variable itself, potentially exacerbating measurement error problems.

Our experiment allows us to compare adjusted R-squared values across three model types: 1). Models with only fixed effects; 2). Models with only temperature proxies; and 3). Models combining proxies with fixed effects. Comparing the first two provides insight into the difference between competing types of proxies in terms of the percentage of variation explained. Comparing the second and third reveals how much of the temperature proxy-induced measurement error is absorbed by the set of fixed effects. Comparing the first and third shows how much of the original variation in the temperature proxy remains after variations associated with the set of fixed effects are purged.

This last comparison has particular empirical relevance since our exercise utilizes high quality U.S. temperature monitors that are widely used in applied research. Typically, outcome variables are functions of both temperature and human influenced activities, meaning fixed effects incorporate both systematic differences across monitoring stations and differences in demographic factors such as age, income and race. In our analysis, the response is simply the true temperature at each location and thus is not confounded with location-specific variables. The percentage of

[17] Although averaging more monitors with distance weights may seem preferable, this is not always true. Along strong temperature gradients (e.g., from the California coast to the inland desert), more distant stations diverge systematically from conditions at the target location. In such settings, nearest-neighbor interpolation can yield smaller errors than weighted averages, because even down-weighted distant stations introduce systematic bias that the weighting scheme cannot eliminate.

[18] Even climate-normal proxies can be improved by adding year indicators, which capture predictable departures from expected daily temperatures driven by phenomena such as the El Niño–La Niña cycle.

true temperature variation explained by different fixed effect sets therefore provides a clean measure of how much original temperature signal remains available for identifying the climate impact coefficient of interest.

Sequential Error Introduction. Having established the idealized natural experiment, we next assess how different sources of temperature measurement error affect estimation of equation (4). We sequentially introduce three error components commonly encountered in applied climate research—monitor interpolation T_{1d}, temporal interpolation T_{2d}, and binning T_{3d} —to evaluate their individual and cumulative impacts on the bias and explanatory power of temperature proxies.[19] For each weather station, we estimate equation (4) using progressively noisier proxies that incorporate these errors in turn. This framework highlights how standard data-construction practices, such as using the nearest available monitor or averaging daily minimum and maximum temperatures, can generate systematic and compounding departures from the true temperature signal. Detailed descriptions of the proxy generation procedures, formulas, and examples are provided in Appendix A.1.

PRISM Proxy. As a benchmark for model-based spatial smoothing, we also evaluate PRISM daily mean temperatures, which integrate both spatial and temporal interpolation within a single modeling framework. PRISM's design allows for comparison against our constructed proxies while also illustrating the influence of complex smoothing algorithms that incorporate topography and climatological data. Details are discussed in Appendix A.2.

3.2 Data Sources and Sample Construction

Our hourly temperature data is taken from the National Centers for Environmental Information (NCEI) Integrated Surface Database (ISD). We obtain hourly temperature data from 2,496 weather stations across the continental U.S. for 2010-2020. To ensure data quality and consistency, we apply two filtering criteria: First, stations with more than 5% missing observations or quality flags indicating potentially erroneous measurements are excluded. This strict threshold helps ensure reliable temperature records across our study period. It also means that a sizable fraction of our retained stations are at airports, which typically maintain high-quality records due to aviation safety and fuel loading requirements. Second, we further restrict our sample to stations with at least one neighboring station within 200km, as our analysis requires paired stations for constructing temperature proxies. After applying these criteria, our final sample consists of 1,039 weather stations. Figure A.1 shows the map of 2,496 active weather monitors

[19]There are three possible sequences in which to consider monitor and temporal interpolation errors: (i) monitor alone, (ii) temporal alone, and (iii) both combined. We report (ii) in Appendix Figure A.7, as it is not required—modern monitors typically report the hourly data needed to avoid temporal interpolation—but it remains of independent interest, particularly when using older observations that report only minimum and maximum temperatures.

during 2010-2020, with colored dots representing 1,039 monitors included in the analysis. The distribution of U.S. weather stations closely mirrors population density, making it particularly useful for socioeconomic and health-related outcome studies.[20]

Our final sample is a panel comprising 4 million observations from 2010-2020. The stations exhibit substantial geographic variation in their characteristics (see monitor summary statistics in Table 1), with elevations ranging from near sea level (0.9 feet) to over 3,000 feet, averaging 395 feet. They are located on average 545km from the coast and 5.7km from the nearest water body. The monitoring network density varies considerably - while stations on average have 19 neighbors within 200km (ranging from 1 to 54 neighbors), the closest neighboring station averages 43km away (ranging from 0 to 185km). Zero-distance between neighboring monitors typically occur when there are monitors at different elevation levels at the same location. All stations in our sample have at least one other station within a 200km search radius, with a maximum of 54 stations near Hampden, MA. For each sample weather station, we also extract the station's PRISM grid cell estimate of daily mean temperature as an alternative proxy measure.

Table 1: Summary Statistics of Sampled Weather Stations

	mean	sd	min	max	unit
Latitude	38.71	5.36	25.48	49.00	
Longitude	-93.93	14.78	-124.56	-67.79	
Elevation	394.86	494.92	0.90	3,027	feet
Dist to coast	545.43	500.23	0.02	1765.22	km
Dist to roads	1.88	2.15	0	44.78	km
Dist to water	5.73	6.14	0.01	63.90	km
Dist to nearest station	43.17	30.73	0	185.02	km
Population density	288.84	1148.21	0	23,516	individuals/km^2
Total stations	18.95	10.39	1	54	
Stations 0-50km	1.58	1.78	0	10	
Stations 50-100km	3.89	3.05	0	16	
Stations 100-150km	5.82	4.04	0	23	
Stations 150-200km	7.66	4.44	0	28	

Note: Latitude and longitude indicate geographic coordinates. Elevation measures height above sea level in feet. Distance metrics include: distance to coast (shortest path to ocean coastline), distance to roads (nearest major roadway), distance to water (closest water body such as lake or river), and distance to nearest station (closest neighboring weather station), all measured in kilometers. Population density represents people per square kilometer in the station's vicinity. Network density is characterized by total station counts within different radii - the total number of stations within 200km and subtotals for 0-50km, 50-100km, 100-150km, and 150-200km bands from each location.

[20] One can draw a subset of U.S. weather monitors relevant to a given activity (e.g., crop acreage), though such networks are necessarily sparser. Unlike pollution monitors, which may be strategically located to avoid non-attainment (e.g., Grainger and Schreiber (2019)), weather monitors are placed where accurate temperature readings are needed (e.g., airports, agricultural stations) and to improve forecasting. Thus, placement mainly affects coverage density and interpolation error, rather than creating selection bias tied to the outcomes of interest.

3.3 Measurement Error Properties

Table 2 reports the summary statistics from the quantitative exercise described in Section 3.1. Panel A reports the summary statistics of the true daily mean temperature $T_{mean,d}$ as well as the proxies that incorporate monitor T_{1d} and temporal interpolation T_{2d} errors.

Panels B and C of Table 2 summarize results from estimating equation (4) separately at each of the 1,039 weather stations for each proxy construction method. Panel B presents the distribution of R-squared values across these station-level regressions, showing that explanatory power consistently falls below 1, with minimum values indicating substantial explanatory power losses where less than 50% of true temperature variance is captured due to accumulated measurement errors. Panel C reveals the distribution of estimated slope coefficients β_1 from equation (4), showing approximately 2% average downward bias for traditional interpolation method (when monitor and temporal interpolation errors are sequentially incorporated) compared to approximately 6% average downward bias for PRISM-based proxies. For traditional interpolation approaches combining both monitor and temporal interpolation errors, estimated slope coefficients across individual monitoring locations exhibit considerable variation, ranging from 0.312 to 1.400.[21] In contrast, PRISM-based proxies demonstrate substantially less variability, with slope coefficients ranging from 0.795 to 1.086, despite exhibiting average downward bias roughly three times larger than traditional interpolation method.

Examining violations of classical measurement error assumptions provides additional insight into how these two proxy types behave. Under classical measurement error, attenuation bias implies that only a small percentage of estimates of β_1 should exceed 1, with this percentage shrinking as the sample size increases under classical measurement error. Our results clearly contradict this expectation. For the traditional nearest-monitor proxy with monitor interpolation error, 44% of ground truth locations yield slope coefficients greater than 1. This proportion decreases to 30% when temporal interpolation error is sequentially added. In contrast, PRISM-based proxies exhibit far fewer violations of the attenuation bias assumption, with only 6% of ground truth locations producing slope coefficients above 1 and a 95% confidence interval for bias ranging from -9.87% to 0.03%. While PRISM exhibits larger absolute biases than nearest-neighbor, its consistently downward bias lends greater credibility to claims that impact estimates are conservative. This finding aligns with Heutel et al. (2021), who document similar attenuation patterns when employing PRISM data for robustness checks.

Figure 1 displays error distributions at three points in each monitor's temperature distribution: the 1st percentile (cold), 50th percentile (median), and 99th percentile (hot). The first three rows show the error distributions from each error source independently, while the bottom row shows their combined effect. The distributions reveal a systematic pattern where

[21] For traditional interpolation, the 95% confidence interval for the slope bias due to monitor interpolation is [-7.61%, 4.92%] while sequentially adding temporal interpolation widens this 95% confidence interval for the bias to [-8.36%, 5.23%].

Table 2: Summary Statistics of Temperature Measurement Error

A. Temperature Proxies

		Mean	SD	Min	Max	N
No Error	T_{mean} (°F)	55.89	19.22	-33.57	114.90	4,128,516
Traditional	+Monitor Intrpl	55.93	19.17	-33.57	111.51	4,128,516
	+Temp Intrpl	56.20	19.33	-32.62	108.50	4,128,516
	+Binning	56.20	19.11	0.00	93.31	4,128,516
PRISM	+PRISM	55.63	19.55	-31.75	109.43	4,128,516
	+Binning	55.63	19.33	2.26	93.40	4,128,516

B. R-squared from Climate Impact Estimations

		Mean	SD	Min	Max	N
No Error		1	0	1	1	1,039
Traditional	+Monitor Intrpl	0.981	0.036	0.523	0.999	1,039
	+Temp Intrpl	0.975	0.036	0.520	0.996	1,039
	+Binning	0.941	0.052	0.392	0.978	1,039
PRISM	+PRISM	0.919	0.028	0.695	0.968	1,039
	+Binning	0.888	0.048	0.441	0.948	1,039

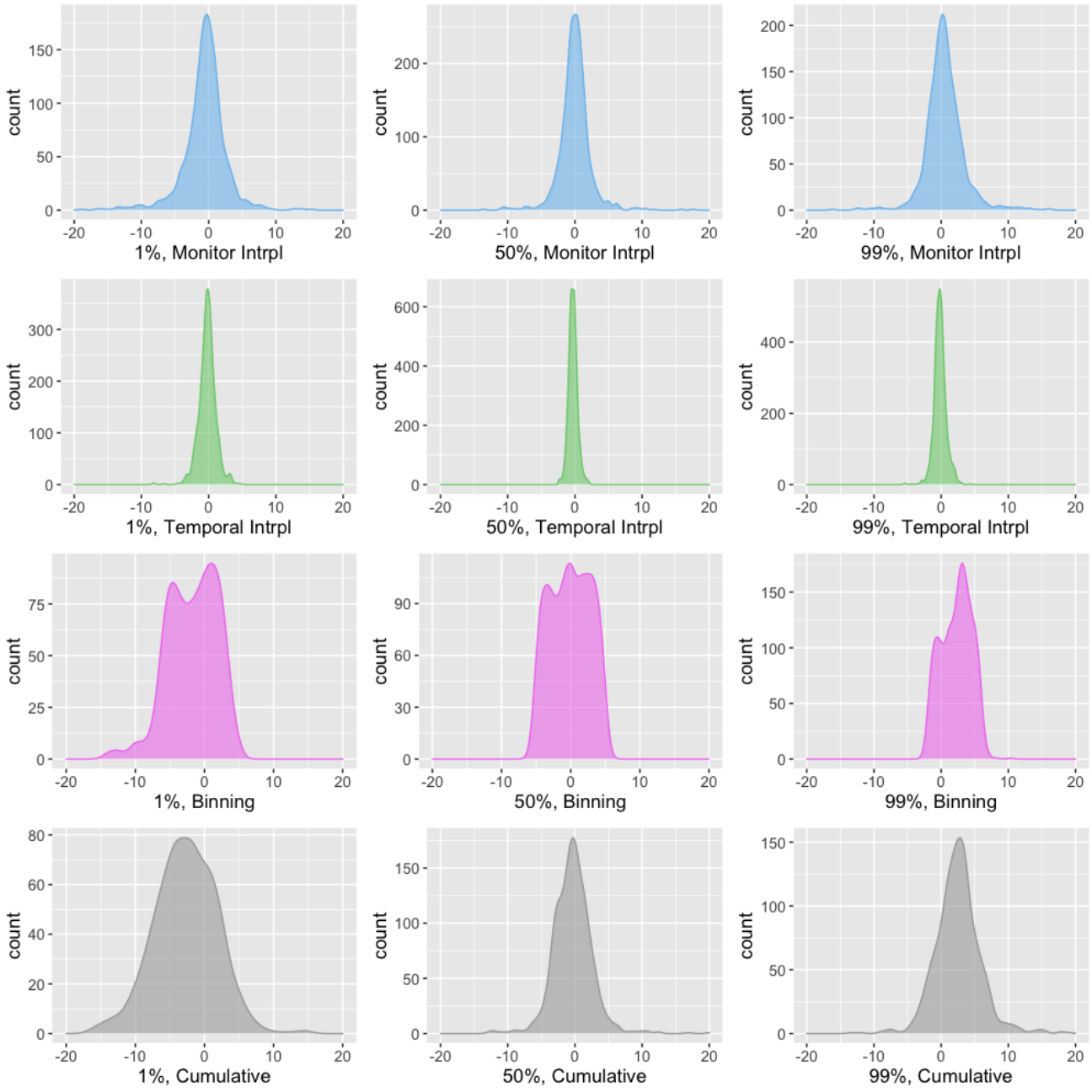
C. Climate Impact Estimations of β_1

		Mean	SD	Min	Max	N
No Error		1	0	1	1	1,039
Traditional	+Monitor Intrpl	0.994	0.054	0.336	1.529	1,039
	$\beta_1 < 1$	0.969	0.047	0.336	0.999	584
	$\beta_1 > 1$	1.026	0.046	1.000	1.529	455
	+Temporal Intrpl	0.984	0.054	0.312	1.400	1,039
	$\beta_1 < 1$	0.966	0.046	0.312	0.999	729
	$\beta_1 > 1$	1.029	0.045	1.000	1.400	310
PRISM	+PRISM	0.942	0.030	0.795	1.086	1,039
	$\beta_1 < 1$	0.938	0.025	0.795	0.999	981
	$\beta_1 > 1$	1.016	0.017	1.000	1.086	58

Note: Panel A reports the summary statistics of temperature proxies constructed using traditional interpolation methods and PRISM. For traditional methods, monitor interpolation (T_{1d}), temporal interpolation (T_{2d}) and binning (T_{3d}) are sequentially added. For PRISM, estimates at our sampled monitoring sites are extracted. Panels B and C report the summary statistics of the R-squared and climate impact estimated from regressing true temperature on these proxies using equation (4) at each of the 1,039 weather stations. Since over- and under-estimations of β_1 are canceled out in the mean, they are reported separately in Panel C.

high temperatures tend to be underestimated while low temperatures tend to be overestimated. This bias arises because both interpolation and binning methods produce smoothing effects that attenuate extreme values toward the mean.

Figure 1: Measurement Errors at 1%, 50%, and 99% of Temperature Distribution



Note: Distribution of measurement error (in °F) displayed at different points in the temperature distribution: the 1st percentile (cold), 50th percentile (median), and 99th percentile (hot) for the 1,039 weather stations. First three rows show the individual contribution of each error source, while bottom row shows combined effect.

We also examine the distribution of PRISM measurement errors across different tempera-

ture ranges in Figure A.2. PRISM errors show greater dispersion than the monitor interpolation and temporal interpolation errors presented in Figure 1, with this pattern especially pronounced at lower temperatures. Because PRISM incorporates readings from our "ground truth" monitors when constructing its temperature estimates, the error estimates in Table 2 likely understate the true magnitude of PRISM measurement errors, particularly in regions with sparse monitoring networks.

The analysis in the rest of this section reveals that these measurement errors exhibit three key patterns. First, they vary systematically with geography. Second, they display strong seasonal patterns. Third, they systematically violate classical measurement error assumptions by exhibiting correlation with true temperatures, location and time. We examine each of these patterns in more detail below.

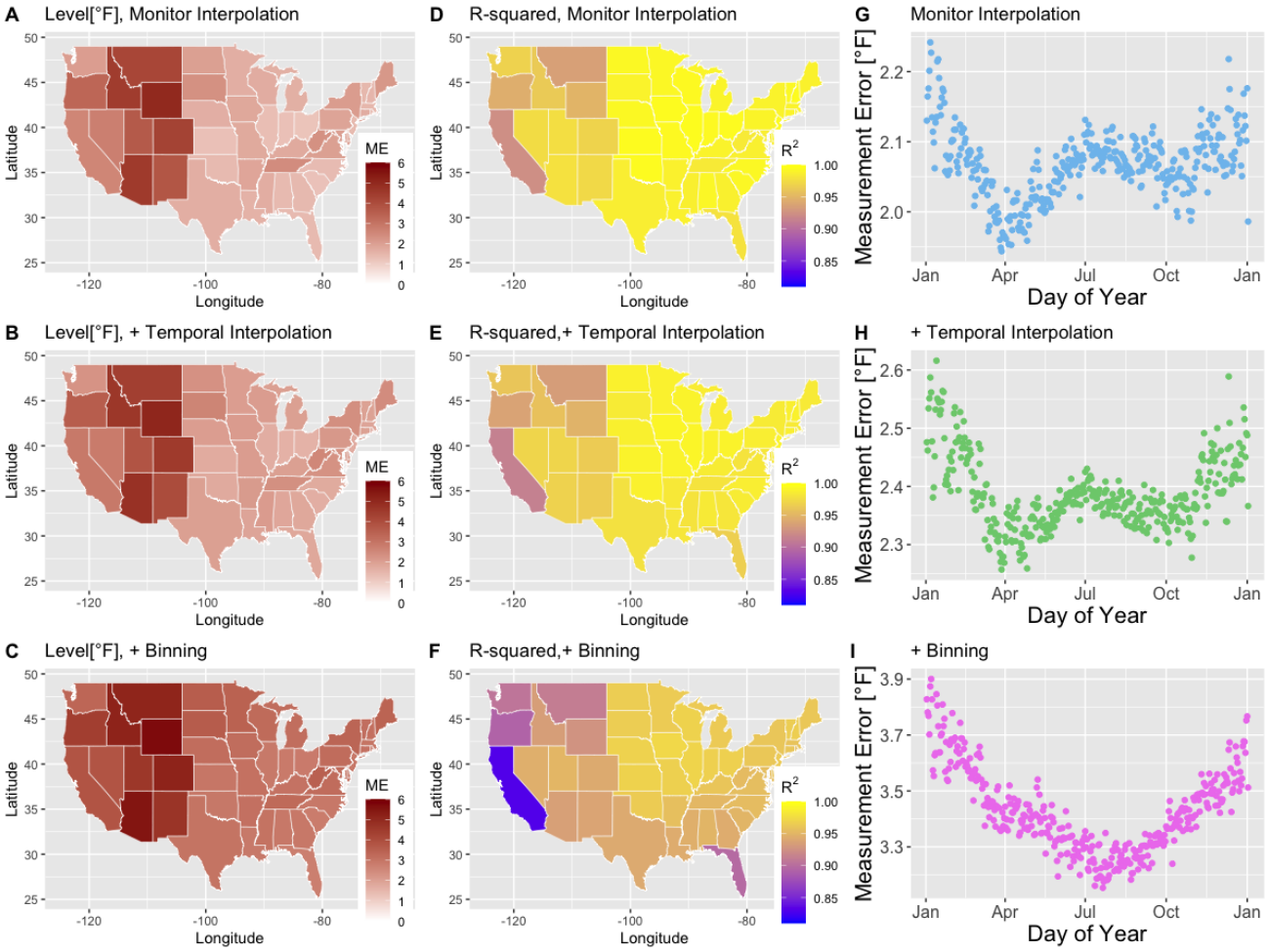
3.3.1 Geographic Patterns

Measurement errors vary systematically with location. As shown in Figure 2(A)-(C), state-level average measurement errors – obtained by averaging daily measurement errors across all stations within each state – are largest in the Rocky Mountain and West Coast states, while errors are smallest in the Southeast and Midwest. This geographic variation stems from fundamental differences in terrain complexity - Western states have complex geographical features coupled with their counties often being larger than counties located more geographically Eastward. The Southeast and Midwest show smaller errors due to their flat topography, relatively small counties, and less variable climates.

These patterns are reflected in the explanatory power of our temperature proxies. Panels (D)-(F) of Figure 2 display state-level R-squared values from equation (4), calculated by averaging station-level R-squared values within each state, indicating how well our proxy measures capture true temperature variation. In California, where terrain varies dramatically, the signal-to-noise ratio ranges from 8:1 down to 2:1 at some weather stations. PRISM shows an average R^2 value of 0.92 compared to ground truth measurements, with more consistent performance across regions than the nearest monitor proxy, but still exhibiting substantive geographic variation.

To explore the determinants of these spatial differences, we relate station-level measurement errors to geographic and network characteristics including elevation (Figure A.5), local monitor density (Figure A.3 and A.4), and land-use heterogeneity (Figure A.6). The results confirm that dense monitoring networks substantially reduce error and that elevation plays a key role, particularly for traditional interpolation methods. The presence of these strong systematic geographic patterns indicate that temperature measurement errors are far from random, violating classical assumptions and highlighting the importance of local conditions for proxy performance. Full regression results are reported in Table A.1.

Figure 2: Measurement Errors Vary with Location and Time



Note: Panels (A)-(C) show state-level average measurement errors in °F for 2010-2020. (D)-(F) show state-level average R-squared estimated from regressing true temperature on proxies (equation (4)). (G)-(I) show daily average measurement errors across 1,039 weather stations. Errors in 2nd and 3rd rows are cumulative.

3.3.2 Temporal Patterns

Measurement errors exhibit strong seasonal patterns. As shown in Figure 2 (G)-(I), when averaged across all weather stations, daily measurement errors are typically larger in winter months and smaller in summer months. Monitor interpolation errors show a unique pattern, reaching their highest levels at both temperature extremes, though this U-shaped pattern becomes less pronounced when combined with other error sources. These systematic seasonal variations provide further evidence that measurement errors violate the independence and identical distribution assumptions of classical measurement error theory.

These seasonal variations help explain the asymmetric distribution of measurement errors we observe. When examining error distributions at different temperature percentiles (see Figure 1), we find that high temperatures tend to be underestimated while low temperatures tend to be overestimated. This pattern is particularly pronounced in winter months when temperature variation within a day is typically larger, making the temporal interpolation error from averaging daily maximum and minimum temperatures more severe. PRISM shows similar seasonal variation in error distributions (Figure A.2), with greater dispersion particularly evident at lower temperatures that occur more frequently in winter months.

The temporal dependence in measurement errors extends beyond seasonal patterns to day-to-day variations. Our analysis of daily errors at representative weather stations suggests strong serial correlation, indicating that measurement errors on consecutive days are not independent. This temporal dependence is particularly evident in the traditional interpolation method, though there is considerable variation in its magnitude across locations.

3.3.3 Single Monitor Tests of the Classical Measurement Error Assumptions

We next examine whether temperature measurement errors satisfy the key assumptions of the classical measurement error framework. Specifically, we test for mean-zero errors, serial independence, correlation with the true temperature, homoskedasticity, symmetric and normally distributed errors. To capture how these properties vary across proxy types, we conduct the tests under five different contexts: monitor interpolation, temporal interpolation, binning, PRISM estimation, and PRISM with binning. The analysis focuses on four representative weather stations—San Francisco (CA), Boulder (CO), Chicago (IL), and Palm Beach (FL)—chosen to reflect a broad range of geographic and climatic conditions.

For each monitor, we compute the measurement error u_{id} as the difference between the proxy and true temperature and apply standard statistical diagnostics. To emulate the moderate sized samples often seen in empirical work, each test is implemented on repeated samples of 500 consecutive days, preserving temporal dependence while yielding more realistic inference.

Table 3 reveals widespread and systematic violations of classical assumptions. Mean-zero errors are frequently rejected, particularly in locations with complex terrain where monitor elevation differences produce persistent bias. Serial correlation is strong in traditional interpolation

Table 3: Test on Classical Measurement Error Properties

	E(u) = 0			Serial Corr.		cov(u,T*) = 0		Hetero.	Skew.	Symm.	Normal.
	(1) Mean	(2) SD	(3) pval	(4) Lag Corr	(5) pval	(6) Corr(u,T^*)	(7) pval	(8) pval	(9) pval	(10) pval	(11) pval
San Francisco, CA											
Monitor Intrpl	0.04	1.68	0.05	0.76	0.00	0.15	0.07	0.00	0.09	0.13	0.02
+Temporal Intrpl	0.81	2.16	0.07	0.65	0.00	0.28	0.02	0.03	0.26	0.22	0.14
+Binning	0.81	2.99	0.08	0.31	0.00	-0.14	0.17	0.03	0.50	0.32	0.07
PRISM	0.57	2.60	0.00	0.10	0.15	-0.04	0.40	0.06	0.10	0.49	0.00
PRISM+Binning	0.57	3.31	0.01	0.16	0.00	-0.32	0.00	0.10	0.33	0.41	0.14
Boulder, CO											
Monitor Intrpl	14.67	5.02	0.00	0.62	0.00	0.43	0.00	0.00	0.00	0.00	0.00
+Temporal Intrpl	15.50	5.00	0.00	0.61	0.00	0.46	0.00	0.00	0.00	0.00	0.00
+Binning	15.50	5.48	0.00	0.46	0.00	0.34	0.00	0.00	0.00	0.04	0.00
PRISM	1.08	6.28	0.01	0.09	0.17	-0.17	0.00	0.00	0.20	0.41	0.00
PRISM+Binning	1.08	6.79	0.01	0.08	0.20	-0.23	0.00	0.00	0.09	0.42	0.03
Chicago, IL											
Monitor Intrpl	-1.36	0.90	0.00	0.44	0.00	-0.07	0.10	0.09	0.08	0.34	0.07
+Temporal Intrpl	-1.34	1.56	0.00	0.17	0.03	-0.01	0.27	0.25	0.20	0.32	0.01
+Binning	-1.34	3.21	0.00	0.06	0.28	-0.13	0.09	0.27	0.22	0.40	0.01
PRISM	-0.52	5.37	0.12	0.01	0.66	-0.08	0.10	0.00	0.32	0.18	0.01
PRISM+Binning	-0.52	5.96	0.14	0.00	0.65	-0.14	0.01	0.01	0.27	0.23	0.06
Palm Beach, FL											
Monitor Intrpl	-2.56	1.38	0.00	0.51	0.00	-0.06	0.14	0.04	0.06	0.39	0.02
+Temporal Intrpl	-2.43	1.94	0.00	0.29	0.00	-0.01	0.22	0.18	0.05	0.27	0.05
+Binning	-2.43	3.31	0.00	0.14	0.06	-0.14	0.09	0.09	0.28	0.44	0.08
PRISM	-0.38	4.38	0.23	-0.03	0.37	-0.06	0.28	0.00	0.23	0.19	0.00
PRISM+Binning	-0.38	5.07	0.27	-0.05	0.34	-0.13	0.01	0.00	0.20	0.50	0.13

Note: Test statistics for classical measurement error properties averaged over 100 random samples of 500 consecutive daily observations at four weather stations. Columns (1)-(3) test whether measurement errors have zero mean, reporting the mean error (in °F), standard deviation, and p-value from t-test. Columns (4)-(5) test for serial correlation using lag-1 autocorrelation coefficient and its p-value. Columns (6)-(7) test for correlation between measurement error and true temperature, reporting the correlation coefficient and p-value. Column (8) reports p-values from Breusch-Pagan test for heteroskedasticity. Column (9) reports p-values from the skewness test for symmetry. Column (10) reports p-values from the new symmetry test. Column (11) reports p-values from the Shapiro–Wilk test for normality.

methods, reflecting day-to-day persistence of local weather patterns, but declines markedly for PRISM proxies that incorporate topographical adjustments. Temporal related heteroskedasticity is also common, with error variance fluctuating seasonally and across months. Most importantly, correlations between measurement error and true temperature are usually statistically significant and often negative, implying that slope estimates can exceed unity—contradicting the attenuation bias expected under classical error. Finally, error distributions are often skewed and clearly non-normal, with long tails.[22]

Together, these results demonstrate that temperature measurement errors are non-classical and structured, not random noise. They vary systematically by geography, proxy construction, and time of year, and thus can bias regression estimates in ways that depend on local conditions and model specification. Details on each test are described in Appendix Section B.2.

3.4 Effect on Climate Impact Estimation

3.4.1 Station Level Analysis

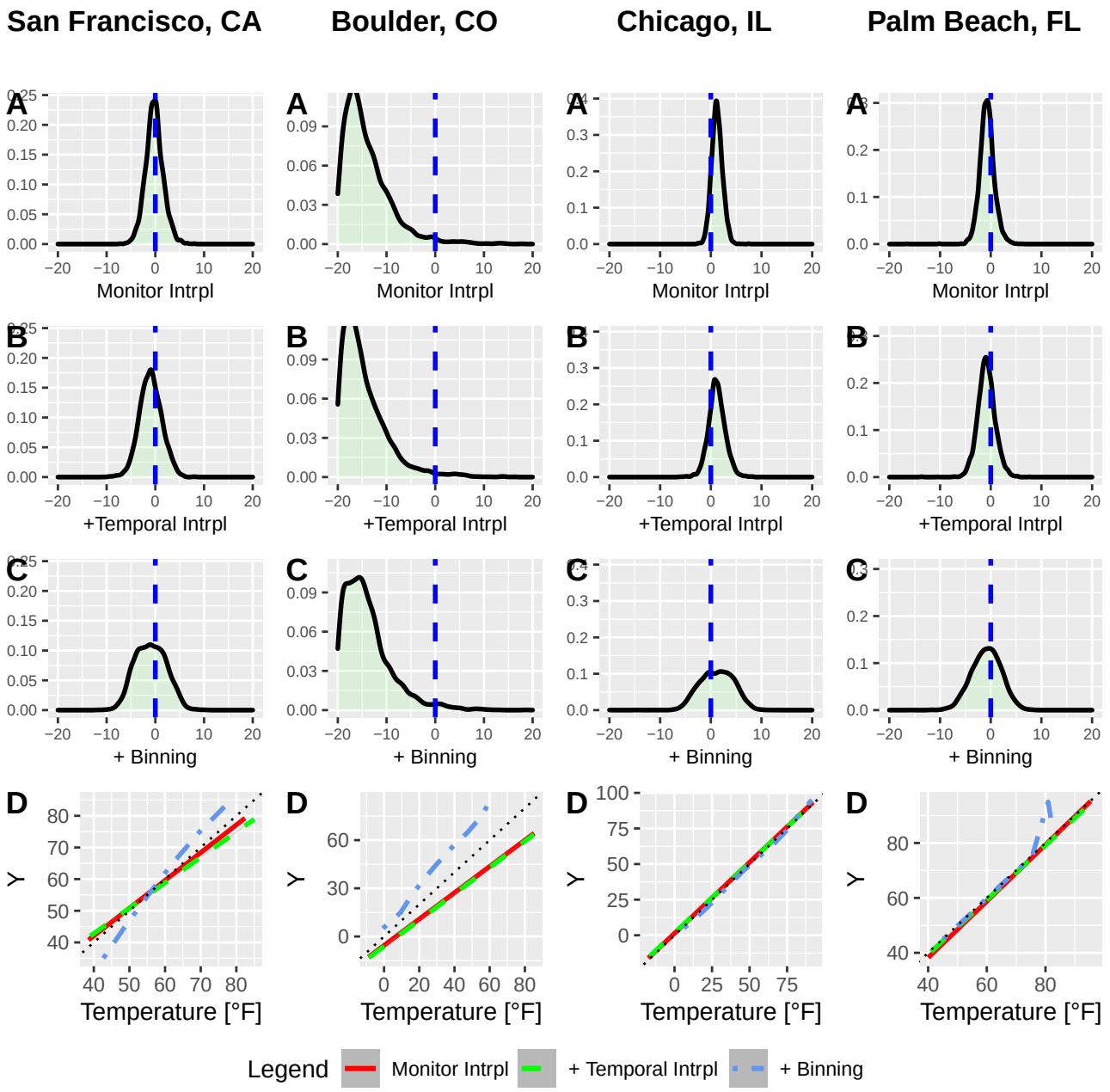
We now analyze how different sources of non-classical measurement error combine to affect climate impact estimates. For each weather station, we estimate variants of equation (4) using the three proxies T_{1d},T_{2d} and T_{3d} as described in Section 3.2. Table 2 Panel C provides summary statistics of these climate impact estimates across all 1,039 weather stations.

The direction and magnitude of measurement bias varies substantially across locations and methods. Figure 3 row (D) demonstrates how the estimated climate impact coefficient β_1 from equation (4) changes with each added error source at the four representative cities. In Chicago, where flat terrain and dense monitoring networks prevail, measurement error effects are minimal. While the conventional wisdom of attenuation bias holds for some locations and error types (as evidenced by the first two slope estimates for San Francisco and Boulder), binning produces overestimation in these cities. San Francisco shows particular sensitivity to different types of measurement error, with slope estimates changing markedly as each error source is introduced. Figure 4 displays these bias patterns, with panels (A)-(C) showing distributions across weather stations and (D)-(F) presenting state-level averages.

We also examine whether estimation bias varies with monitoring network characteristics in Table A.1. For traditional interpolation, we find minimal correlation between estimation bias and key network metrics – correlation coefficients of just -0.03 with the number of nearby stations within 200km and 0.08 with distance to the closest station. Similarly, while PRISM-based estimation bias shows slightly stronger correlations with network characteristics (-0.17 with nearby station count and 0.23 with distance to nearest station), these relationships are still relatively modest.

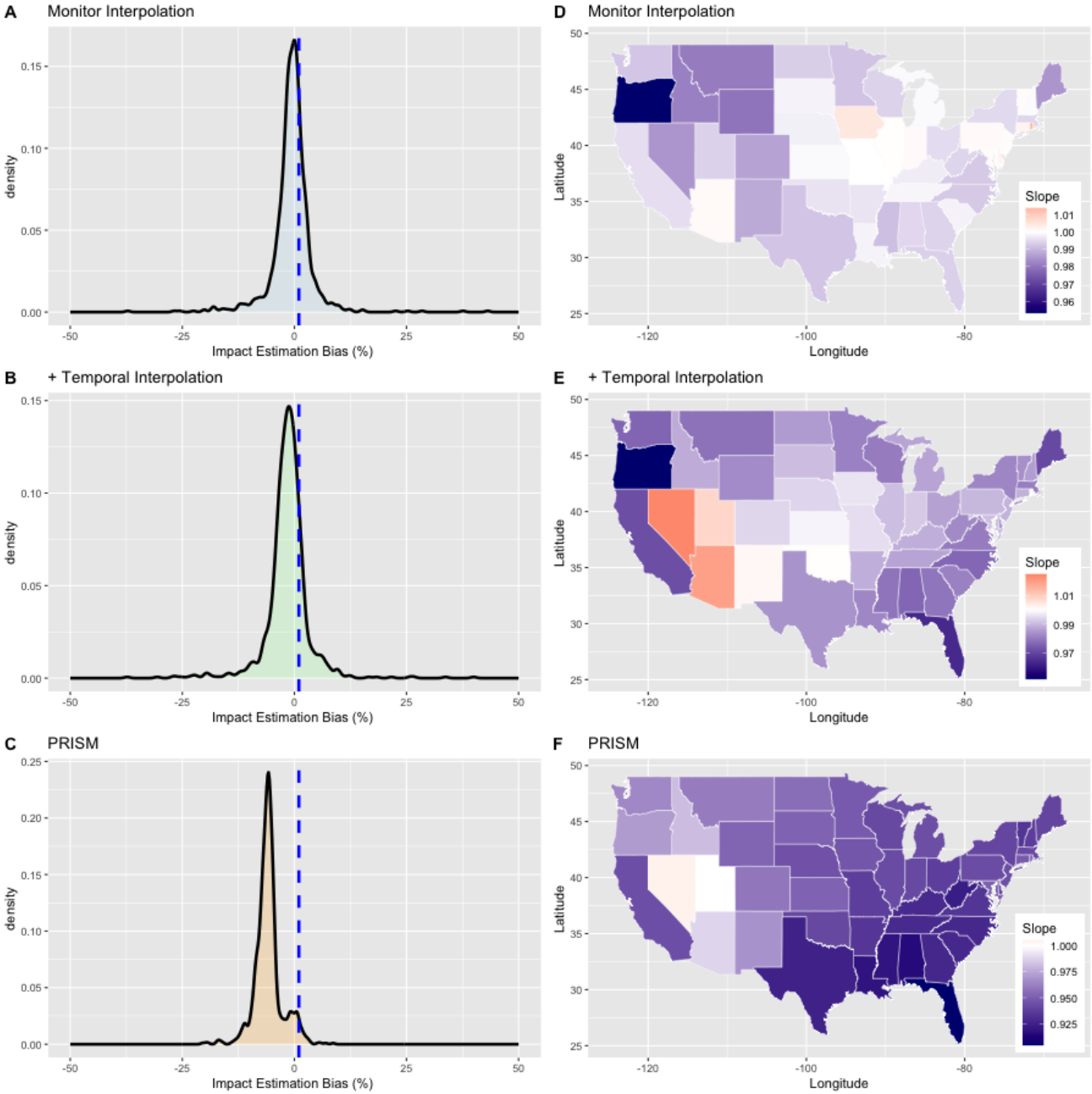
[22]Such skewness parallels recent work on one-sided measurement error (Millimet and Parmeter, 2022), which develops methods to correct bias from skewed error distributions. Their framework focuses on outcome variables but offers useful intuition for addressing asymmetric error in covariates.

Figure 3: Measurement Errors' Impact at Four Representative Cities



Note: Based on 2010-2020 data for weather monitors in the four representative cities. Rows (A)-(C) show the distribution of cumulative measurement errors after each error is added. Row (D) shows the slope coefficient β_1 estimated from equation (4). A line with a slope above the 45 degree line represents an overestimate of the climate impact and below the 45 degree line an underestimate.

Figure 4: Climate Impact Estimation: Average Impact at Each State



Note: Impact of measurement error on climate impact estimation estimated with equation (4) using two proxies: traditional interpolations and PRISM data. Panels (A)-(C) display distribution of bias when estimating slope coefficient from equation (4). Panels (D)-(F) display state-level average β_1 parameters estimated with equation (4). $\beta_1 < 1$ indicates an underestimation of the climate impact and $\beta_1 > 1$ an overestimation.

3.4.2 Aggregated Analysis

Since many economic studies of climate impacts examine broad geographical regions rather than single locations, we extend our analysis of measurement error to a larger spatial scale. To do this, we pool observations across all weather stations into a single regression framework. Specifically, we stack the daily "true" mean temperature observations T_{mean} into a vector and regressing them on our constructed proxy measures. This pooled analysis focuses on monitor interpolation and temporal interpolation errors, where ground truth comparisons are well-defined.

The baseline OLS regression yields:

$$\begin{aligned} \widehat{T}_{mean,id} &= 0.875 + 0.979 \times \text{Proxy}_{id} \\ &\quad (0.006) \quad (0.000) \qquad\qquad \text{Adj.}R^2 = 0.970 \end{aligned}$$

Here, Proxy_{id} represents the proxy temperature at monitor i on day d. This proxy is constructed using the closest neighboring station's daily mean temperature, calculated as the average of the minimum and maximum temperatures recorded at that station.[23]

This regression yields an adjusted R^2 of 0.970 and RMSE of 3.336°F. Under the null hypothesis of no measurement error, we would expect an intercept of 0 and both slope coefficient and R^2 of 1. The observed slope coefficient of 0.979 indicates modest downward bias. The non-zero intercept of 0.875 means our proxy underestimates true temperature at 0°F by 0.875°F, while overestimating it by 1.25°F (calculated as (100-0.875)/0.979-100) at 100°F and by 1.47°F (calculated as (110-0.875)/0.979-110) at 110°F.

The policy relevance of these temperature proxy measurement errors is obviously context dependent. Since we have explicitly cast the response variable in terms of temperature, a natural comparison point is temperature predictions from climate models. Here the measurement errors found are substantial – translating roughly into an increase in the climate model's prediction errors of ±1.2-1.4°F at higher temperatures. The RMSE of 3.336°F (1.853°C) of the measurement error exceeds the Paris Agreement's 1.5°C temperature increase target by 24%.

3.4.3 Aggregated Analysis with Fixed Effects

The same stacked regression approach is used to examine what fraction of temperature variation is explained by different sets of commonly-used fixed effects. Table 4 panel A reports the adjusted R^2 for models with seven different sets of time and location fixed effects.[24] The base case with no fixed effects yields an adjusted R^2 of 0. Column (2) shows that using a set of monthly fixed effects explains 57.4% of the variation in daily mean temperature at our ground truth monitor

[23] This proxy measure corresponds to T_{2d} in the station level analysis in Section 3.1. The key difference here is the inclusion of monitor subscript i, as this is a pooled regression across multiple monitors

[24] Our fixed effect sets can be implemented in a statistically equivalent manner by a set of indicators in a regression model where the coefficient on one of the indicators is set to zero, or where the indicators are defined in terms of deviations from the overall grand mean, which is common with ANOVA approaches.

sites. Moving from monthly to day-of-year fixed effects (column 3) adds over 300 coefficients but increases adjusted R^2 by only 1.6%. Location-based fixed effects (columns 5-7) explain substantially less temperature variation: state-level fixed effects explain 18.2%, county-level 21.1%, and monitor-level 21.4%. Note that a monitor-level fixed effect is empirically equivalent to including a unit-specific (e.g., individual) fixed effect. Of course, it is common in applied work to use both time and location fixed effect sets and their cross products, as shown in columns (4) and (8) with the fixed effects alone absorbing approximately 80% of variation in the true temperature.

In Panel B, we add these different fixed effect sets to our standard model that uses the nearest monitor's daily mean temperature as the proxy. Most fixed effect sets modestly improve adjusted R^2 relative to baseline (column 1). Monitor and county-specific fixed effects show larger improvements, though at the cost of adding over 600 and 1,000 degrees of freedom, respectively. Use of time-based fixed effects notably reduces the proxy slope coefficient. In the base model (column 1), the proxy slope coefficient indicates 2.1% downward bias from the true value of 1. Inclusion of monthly fixed effect in the column (2) model more than doubled this bias to 4.7%.

Panel C shows that PRISM-based proxies exhibit larger baseline bias (4.7% versus 2.1%), equivalent to the Panel B column 2 model with monthly fixed effects. As in Panel B, adding sets of time-based fixed effects increases the bias in the slope coefficient. Here, introduction of monthly fixed effects in column (2) increases the bias in the slope coefficient to 8.7%. Turning to location-specific fixed effects, the increase in bias is not as large as for time-based effect sets, but it is always larger than in the base (column 1) model. The popular state $\times$ month fixed effects (column 8) increases the bias to 15.3% , more than tripling the baseline bias. Since these fixed effects explain 78.8% of true temperature variation, only 21.2% remains for parameter identification. Furthermore, the inclusion of proxy variables based on temperature readings from other monitors, whether using our nearest-neighbor approach or PRISM, introduces additional measurement error that further reduces this identifying variation.

To examine heterogeneous regional patterns, we estimate separate regressions for each U.S. region (Midwest, Northeast, South, and West). Figure 5 reveals that PRISM performs better in western states, while our nearest monitor proxy dominates in other regions.

4 Spatial and Temporal Aggregation: Weather and Crime

Now we turn to the question of how climate impact estimates are influenced when measurement errors due to spatial and temporal aggregation are introduced. In doing so, we move beyond our previous framework which was focused on specific point locations defined by latitude and longitude at the daily level. In many instances, the data available to researchers has been aggregated either across space, time or both. To address how these aggregation processes shape estimated effects, we examine the extent to which measurement error in temperature alters the

Table 4: Temperature Measurement Error Regression Results

A. Fixed Effects Only

Model:	Base	Month	DOY	DOY× Monitor	Monitor	County	State	State× Month
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Adj R^2	0.000	0.574	0.590	0.840	0.214	0.211	0.182	0.788

B. Traditional Interpolation Methods

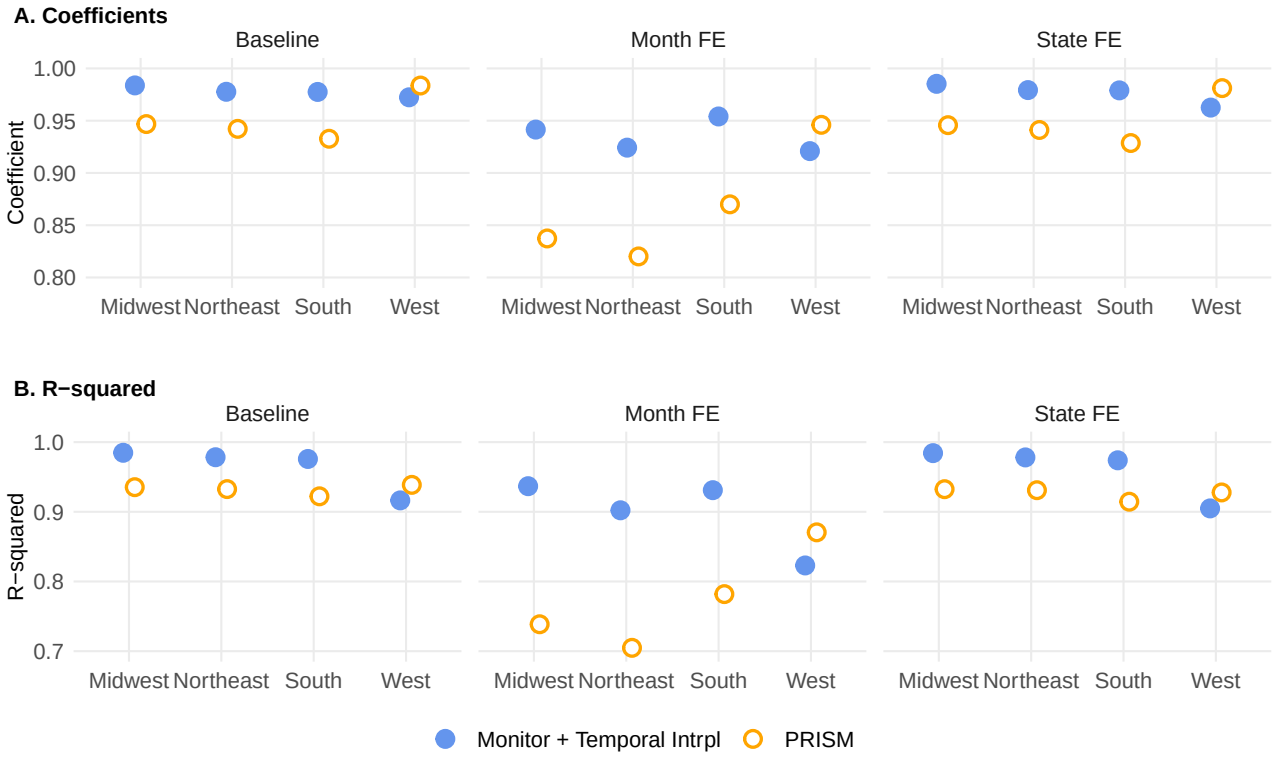
Model:	Base	Month	DOY	DOY× Monitor	Monitor	County	State	State× Month
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	0.875***							
	(0.006)							
Proxy	0.979***	0.953***	0.951***	0.943***	0.987***	0.985***	0.978***	0.921***
	(0.0000)	(0.0001)	(0.0001)	(0.0002)	(0.0000)	(0.0000)	(0.0000)	(0.0002)
Adj R^2	0.970	0.970	0.971	0.986	0.984	0.980	0.971	0.971
Within R^2		0.931	0.928	0.912	0.979	0.975	0.964	0.867

C. PRISM

Model:	Base	Month	DOY	DOY× Monitor	Monitor	County	State	State× Month
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	2.886***							
	(0.008)							
Proxy	0.953***	0.913***	0.911***	0.798***	0.945***	0.946***	0.948***	0.847***
	(0.0001)	(0.0002)	(0.0002)	(0.0003)	(0.0001)	(0.0001)	(0.0001)	(0.0003)
Adj R^2	0.940	0.941	0.942	0.946	0.941	0.941	0.940	0.943
Within R^2		0.862	0.858	0.661	0.925	0.925	0.927	0.732

Note: Regression results using different sets of fixed effects specifications. Panel A shows results using fixed effects only, Panel B shows results using traditional interpolation methods, while Panel C shows results using PRISM temperature data. Dependent variable is the true daily mean temperature from weather station observations. Proxy refers to either the traditional interpolated temperature (Panel B) or PRISM temperature estimate (Panel C). Models include various fixed effects: month, day-of-year (DOY), DOY-by-monitor interactions, monitor locations, counties, states, and interactions of geographic units with months. Within R-squared shows explanatory power after removing fixed effects. Robust standard errors are in parentheses. The sample includes 4,128,516 daily temperature observations from 1,039 weather stations over 2010-2020.

Figure 5: Heterogeneity in Impact Estimation: by Region



Note: Regression coefficients (top panels) and R-squared values (bottom panels) from models predicting true temperature using two measurement approaches across U.S. regions, estimated by running equation (4) separately for each of the four U.S. regions (Midwest, Northeast, South, and West). Blue dots/triangles represent traditional interpolation method (Monitor + Temporal Interpolation) while orange dots/triangles represent PRISM estimates. Results are shown for three model specifications: baseline model, models with month fixed effects, and models with state fixed effects. The coefficients indicate the relationship between measured and true temperatures (with 1.00 representing perfect prediction), while R-squared values show the proportion of variation in true temperature explained by each proxy measure. For models with fixed effects, "within" R-squared – the proportion of variation explained by the measurement approaches after removing the variation explained by the fixed effects is reported.

often studied relationship between weather and crime in the United States, where data may be available at different spatial scales (e.g., ZIP code, city, county, or state) and over different temporal resolutions (e.g., hourly, daily, monthly, or annual).

This application is not intended as a replication or critique of any particular study in the weather–crime literature. Rather, we use crime as a well-studied setting in which temperature effects are known to be nonlinear and economically meaningful, making it a useful benchmark for assessing how aggregation-induced measurement error propagates through standard empirical designs. Our baseline specification follows the widely used binned temperature framework, similar in structure to Heilmann et al. (2021), although differences in geography and temperature range preclude direct replication. At the same time, the overall shape of our estimated response function is broadly consistent with nationwide evidence such as Berman et al. (2020), despite differences in modeling strategy. Finally, although daily data are available for crime, we also report monthly specifications to illustrate how temporal aggregation—often unavoidable in other outcome settings such as health or mortality—can smooth nonlinear responses that are visible at higher frequencies.

4.1 Crime Data and Methods

Our empirical analysis starts with a daily panel of violent and property crime constructed from incident-level reports submitted by participating law enforcement agencies to the FBI's National Incident-Based Reporting System (NIBRS) and companion state reporting systems. These reports are aggregated to the city–day level and cover more than 5,000 cities in the contiguous United States during the 2010–2019 period. Violent crime consists of offenses involving force or the threat of force, including aggravated assault, robbery, rape, and homicide. Property crime includes burglary, larceny, motor vehicle theft, credit card theft, and other non-violent offenses involving the taking or destruction of property.

For our baseline temperature measure, we use inverse distance weighted (IDW) averaging of monitor readings at city population centroids. This approach is commonly used when the outcome variable is defined at an aggregated spatial level. For our purposes, we assume the IDW-based measure represents the best available approximation to the temperature conditions that residents of each city experience on a given day. Let T^*_{id} denote this assumed "true" daily maximum temperature in city i and day d.[25] We then estimate variants of the following specification:

$$\text{Crime}_{id} = \sum_{j=1}^{10} \beta_j \text{TBIN}_{j,id} + \delta \text{PRCP}_{id} + \mathbf{X}_{id}\boldsymbol{\eta} + \epsilon_{id} \tag{5}$$

where Crime_{id} denotes the the number of violent or property crimes per 100,000 people in city

[25] In the appendix, we also report results using daily mean temperature constructed with the monitor and temporal interpolation errors examined in Section 3.

i on day d. $\text{TBIN}_{j,id}$ are a set of indicator variables equal to one when the true temperature T^*_{id} falls into one of the ten temperature bins (below 10°F, 10-20°F, 20-30°F, 30-40°F, 40-50°F, 50-60°F, 60-70°F, 70-80°F, 80-90°F, and above 90°F), with the 60-70°F bin serving as the reference category. Daily precipitation is captured by PRCP_{id}, which is drawn from PRISM data and included uniformly across all specifications for consistency. The vector $\mathbf{X}_{id}$ includes city fixed effects, year–month fixed effects, and day-of-week indicators.[26] All regressions are weighted by city population, and standard errors are clustered at the city level.[27]

We then evaluate several commonly used temperature aggregation schemes that capture the major sources of spatial and temporal measurement error. To assess spatial aggregation, we move from city–day exposure to county–day exposure by constructing county-level temperature bins from the IDW-based measure T^*_{id}. This step captures the additional measurement error introduced when within-county spatial heterogeneity is collapsed to a single county-wide value.

Turning to temporal aggregation, we replace daily temperatures with county–month exposures. For each county–month, we count the number of days falling into each temperature bin and aggregate crime outcomes accordingly. This procedure potentially smooths short-run variation, especially extreme temperature days that are central to the temperature–crime relationship. Throughout all spatial and temporal aggregation exercises, we ensure that the coefficients on each temperature-bin indicator retain a consistent interpretation: one additional day in bin TBIN_j is associated with β_j additional crimes per 100,000 people.

We also construct PRISM-based proxies by extracting predicted temperatures at county centroids. This provides a model-based counterpart to the monitor-derived measures but introduces its own smoothing-related error. Because temperature enters through discrete bins, proxy-specific measurement error affects bin assignment, so differences across proxies partly arise from misclassification induced by the interaction between measurement error and binning.

In addition to our baseline use of daily maximum temperature, the most commonly used temperature stimulus, we also construct an analogous set of results using daily mean temperature in Table A.4. Daily mean temperature allows us to directly incorporate the monitor interpolation and temporal interpolation errors documented in Section 3. This alternative specification therefore provides a comprehensive assessment of all measurement-error channels discussed in the paper—including nearest-neighbor substitution, averaging of daily minima and maxima, spatial and temporal aggregation, and binning.

[26] Similar to the specification in Heilmann et al. (2021), which investigates the temperature–crime relationship in Los Angeles. Their main specification uses two high-temperature indicators (75–89°F and ≥90°F) relative to cooler days, reflecting the narrower temperature range in a single-city setting.

[27] For county-level estimations, we analogously include county fixed effects, weight by county population, and cluster standard errors at the county level.

4.2 Results

The estimated temperature–crime relationships from the city–day data using monitor-based temperatures serve as our benchmark against which all aggregation-induced distortions are evaluated and are reported in Table A.3 and Figure 6. As shown in the black curves in Figure 6 Panel A, property crime is substantially lower on cold days, rises steadily through moderate temperatures, and peaks on the hottest days (the sample mean is 16.22 per 100,000). Relative to the 60–70°F baseline, the <10°F coefficient is –1.62 per 100,000 people, and the 90°F+ coefficient is 0.28, consistent with prior work on temperature–crime sensitivities (Berman et al., 2020; Heilmann et al., 2021).[28]

Introducing spatial aggregation by moving to county–day temperatures yields a systematic shift in the response function. Because within-county temperature heterogeneity is averaged into a single exposure measure, the usable signal in temperature falls relative to the noise arising from geographic smoothing. The resulting estimates, shown in Panel A by the orange line, are consistently closer to zero at both the cold and hot ends. For example, comparing the IDW estimates (Table A.3), the <10°F coefficient increases 30% from –1.62 to –1.29, and the >90°F coefficient declines 25% from 0.28 to 0.21. The NN estimates are marginally larger than the IDW estimates and the PRISM estimates are somewhat larger than the NN estimates. These deviations occur because cold days are disproportionately misclassified upward into warmer bins, and hot days are averaged with milder days, reducing within-bin variation and narrowing the estimated gradients.

Temporal aggregation has far more pronounced consequences. When temperatures are aggregated to the county–month level, the estimated gradient steepens sharply at the ends of the distribution. In Figure 6 Panel B, the coefficients for the coldest bins become roughly twice as large in magnitude as their city–day counterparts; the <10°F estimate falls to –2.61 under monitor data and –2.44 under PRISM (Panel D). At the hot end, the differences are even more striking: the >90°F coefficient rises from 0.28 in the city–day benchmark to 0.86–0.93 in the county–month specifications, indicating biases of greater than 300% relative to the daily IDW daily city level estimates. These changes reflect the loss of intra-month variation and the inherently nonlinear temperature–crime relationship, which together cause the month-level measure to overweight extreme-temperature days relative to more moderate periods. Violent crime exhibits analogous patterns (Panels E–H), although with smaller percentage bias magnitudes. The sharp change in relative magnitudes between the 10–20°F and 20–30°F bins reflects this same nonlinearity: as the daily response function flattens at low temperatures, smoothing across adjacent bins can attenuate the coldest effects while exaggerating the neighboring bin.

The daily-mean temperature results (Table A.4) are quite similar to those using daily

[28] Our high-temperature effects are directly comparable to those in Heilmann et al. (2021), which find that in Los Angeles, days above 90°F increase reported crime by about 1.9%. Converting their estimate to our units implies an increase of roughly 0.2 to 0.3 crimes per 100,000 people, which is very similar to our city–day estimates.

maximum temperature, except for an upward shift in bin density at the low end and a downward shift at the high end. This is not surprising, as daily maximum and mean temperature proxies are highly correlated (0.98 in the daily IDW city-level data). Overall, the results exhibit the same qualitative distortions as the daily-maximum specification, confirming that the aggregation-induced biases we document are robust to the choice of temperature measure and to the inclusion of all major measurement-error channels.

Taken together, this section shows that aggregation-induced measurement error does not simply attenuate temperature effects toward zero. Rather, the direction and size of the biases vary systematically across the distribution, with spatial aggregation reducing curvature and temporal aggregation amplifying it, especially at the tails.

5 Discussion

5.1 Guidance for Empirical Researchers

The empirical findings presented in the preceding analysis provide a foundation for several actionable recommendations for researchers working with temperature data. While our empirical analyses focus on specific interpolation and aggregation methods, the underlying insights are general and extend to many settings where environmental data must be constructed from imperfect proxies. This subsection consolidates those lessons into a set of guiding principles concerning proxy selection, data construction, and model specification.

5.1.1 Choosing Among Temperature Proxies

The choice of interpolation method should reflect a fundamental tradeoff between bias and variance. Researchers often rely on two widely used temperature proxies: readings from the nearest ground-based monitor (traditional interpolation) and estimates from gridded data products such as PRISM.[29] A key finding from our analysis is that sophisticated gridded products do not universally or even generally outperform simpler approaches. In fact, when monitoring networks are sufficiently dense—as in many urban settings—nearest-neighbor interpolation often produces smaller measurement error and more accurate temperature estimates.

Traditional interpolation using the nearest monitor tends to yield impact estimates with relatively small downward bias of 3%. However, biases associated with this approach vary considerably across space and can occur in either direction, from roughly 70% too small to 40% too large. This wide dispersion implies that studies conducted in specific locations may estimate effect sizes that differ markedly from their true values.

[29]PRISM is one of several widely used gridded temperature products. Other prominent examples include Berkeley Earth, MODIS, the North American Regional Reanalysis (NARR), and the Weather Research and Forecasting (WRF) model. The properties of proxies based on such products can be compared using the approach put forward in this paper.

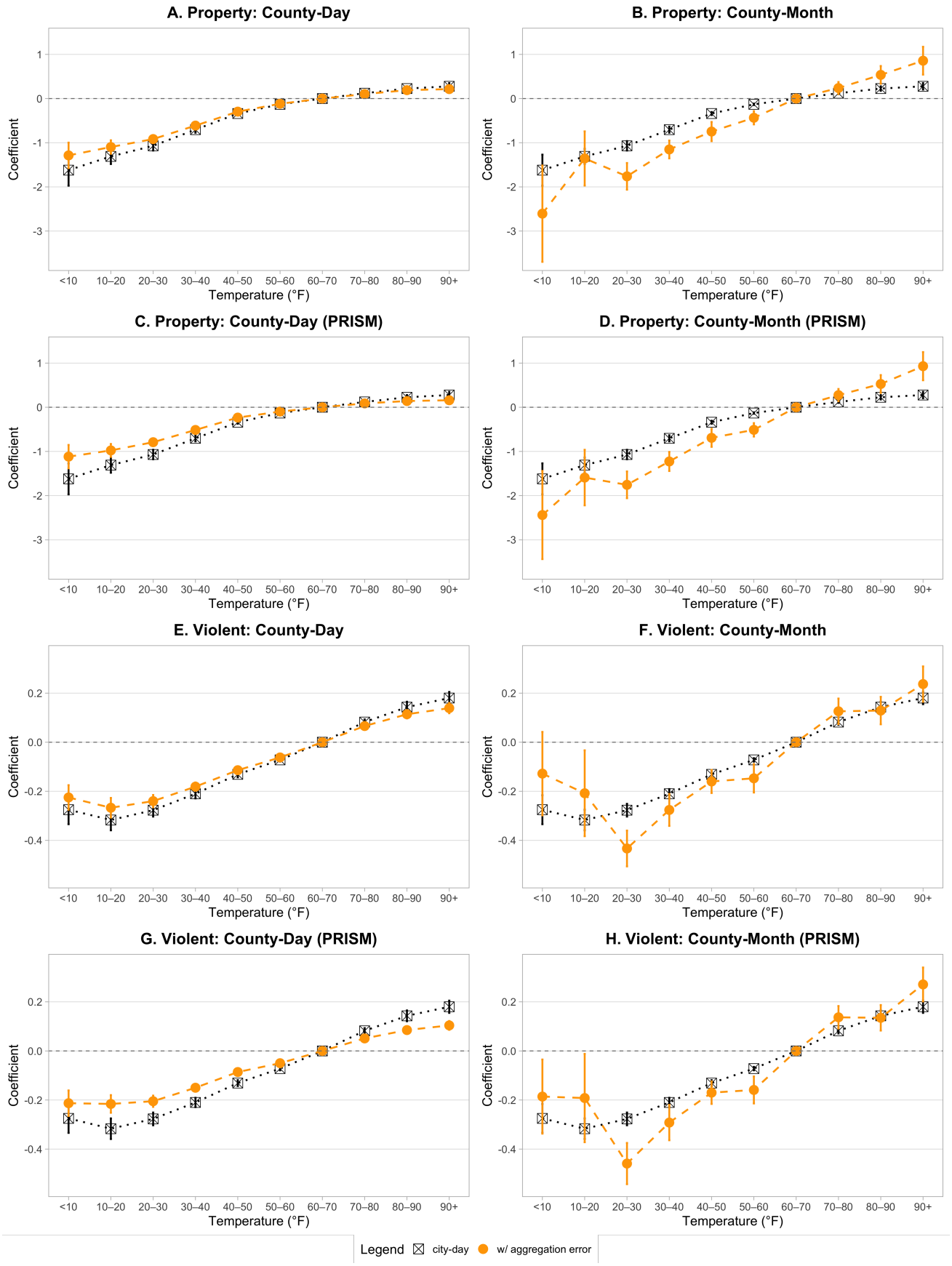


Figure 6: Estimated Relationship between Temperature and Crime

Notes: Temperature–crime estimates using city–day level data (square) versus county–day/county–month level data (circle). Each point shows estimated effect of a day in each temperature bin on crime counts relative to a day in the 60–70°F bin. Regressions are weighted by population and include controls for precipitation, city/county fixed effects, day-of-week indicators, and state-by-month fixed effects. Error bars represent one standard deviation above and below the point estimates. Standard errors are clustered at the city or county level.

In contrast, PRISM-based proxies exhibit the opposite pattern. Their smoothing algorithms generate larger absolute biases of 6% on average, but PRISM's measurement errors are more likely to take the form of attenuation bias, rather than the bidirectional errors observed in traditional interpolation. Further, the magnitude of this downward bias increases considerably with commonly used fixed-effect specifications. This pattern makes PRISM particularly suitable when researchers desire conservative and directionally consistent estimates, even if those estimates are somewhat muted in magnitude.

Ultimately, the relative performance of these proxies depends on geographic and data contexts. Ground-based nearest-neighbor interpolation is typically preferable in low-elevation regions with dense monitoring coverage, where it captures fine-scale local variation efficiently. In contrast, PRISM performs better in areas with complex terrain or sparse monitor networks, where its spatial smoothing reduces idiosyncratic noise and yields more stable, interpretable relationships. Together, these results emphasize that the appropriate proxy choice depends not only on data availability but also on researchers' tolerance for variance versus directional bias.

5.1.2 Constructing and Aggregating Temperature Data

Constructing temperature data carefully can substantially reduce measurement error. When feasible, daily mean temperature should be computed as the average of 24 hourly observations, which accurately captures the full diurnal cycle. Such hourly data has long been available from most high quality weather monitors. The common practice of averaging daily minimum and maximum temperatures should be avoided unless a very long time series is required, in which case some type of seasonal correction may be needed to mitigate bias.

A more reliable approach to spatial aggregation is to construct a temperature exposure measure that reflects as closely as possible where relevant activity actually occurred. This can be implemented either by weighting proxy temperature measurements at the centroids of small geographic units (such as U.S. Census block groups) according to the density of treatment units, or by aggregating temperatures from gridded data products using analogous activity-based weights. For socioeconomic outcomes, temperature exposure should typically be population-weighted (e.g., Auffhammer, 2022; Baylis, 2020; Burke et al., 2018; Carleton et al., 2022). For agricultural outcomes, weighting by cropland area is more appropriate (e.g., Schlenker and Roberts, 2009).

Finally, as shown in our empirical application, reducing the size of spatial units helps limit ecological-fallacy risks (Robinson, 1950). However, reducing the degree of temporal aggregation in our empirical example is even more important for preserving meaningful variation in temperature exposure—particularly when responses are nonlinear or concentrated in the tails of the temperature distribution. Whether the greater sensitivity to temporal rather than spatial aggregation is application-specific remains an open question. It is also clear that the operational definition of the temperature stimulus becomes increasingly important over longer time horizons,

such as in agricultural production studies, where exposure to the same daily temperature may be beneficial during some lifecycle stages and detrimental during others. Transparent reporting of these design choices can help clarify the likely direction of measurement error and improve comparability across studies for meta-analyses.

5.1.3 Choosing Model Specification

Even after minimizing measurement error in temperature proxies, specification choices such as fixed effects and nonlinear forms can substantially influence how that error propagates into estimated effects. Fixed effects are standard practice for controlling unobserved stimulus and would otherwise bias the estimated temperature responsiveness. However, their use can substantially deplete the signal in temperature measures because fixed effects mechanically absorb variation that helps predict true temperature. In the absence of measurement error, this simply reflects the usual trade-off: fixed effects yield consistent estimates but at the cost of reduced precision. When temperature is measured with error, the trade-off becomes more problematic. Measurement error not only reduces precision but also biases the estimated temperature effect, and this bias grows as measurement error becomes large relative to the remaining true variation after fixed effects are applied. As such, applied researchers need to carefully weigh the trade-off between controlling for unobserved heterogeneity and amplifying measurement error bias. Some insights into this tradeoffs can be gained by examining how temperature responsiveness changes with different fixed effects specifications.

Robustness can be assessed by varying the fixed effect specification – sequentially adding and removing sets of fixed effects – to reveal how coefficient magnitudes and precision change, making the trade-off between confounder control and measurement error bias transparent. This also emphasizes a deeper point: every fixed effect specification implicitly defines a different expected temperature profile, and the estimated temperature effect is identified from deviations around that expectation. Recognizing this can help clarify what, exactly, each fixed effect is intended to control for.

Researchers should also exercise caution when employing nonlinear functional forms, such as the widely used binned regression approach. Nonlinear models offer greater flexibility, particularly for detecting non-linear responses at temperature extremes where climate impacts are often thought to be concentrated. Yet measurement error is also largest at these extremes: high temperatures tend to be underestimated and low temperatures overestimated. In binned regressions, this pattern produces systematic misclassification across temperature bins, and the misclassification becomes increasingly one-sided in the tails. This asymmetric misclassification pattern can distort the estimated temperature-response curve precisely in the ranges of greatest interest. Although our discussion focuses on binned regression models, other flexible specifications that give tail observations disproportionate influence, such as polynomial or spline-based approaches, are similarly vulnerable to distortion in the presence of measurement error. A com-

parative assessment of how different flexible functional forms respond to the types of errors documented here would be a valuable direction for future work.

5.2 Prospective Uses of Framework

The analyses in this paper provide a systematic characterization of how temperature measurement error behaves in practice. The errors we document are neither small nor random; instead, they exhibit persistent spatial and temporal structure that challenges the classical assumptions underpinning most empirical models. Measurement errors are correlated across time and space, concentrated in certain regions and seasons, and frequently large relative to the variation researchers seek to explain. Importantly, these errors are not symmetrically distributed, so that the expected attenuation bias of classical measurement error theory does not necessarily hold. In many contexts, the resulting bias can either overstate or understate true effects, depending on how proxies distort the local temperature distribution.

Beyond documenting these empirical regularities, the framework developed here provides a generalizable tool for diagnosing and quantifying measurement error in other environmental data. By comparing proxy-based estimates to an observed "ground truth" the framework isolates the structure and magnitude of measurement error in a way that is both transparent and replicable. It is possible to draw from the location-specific measurement error distributions we identify to conduct robustness exercises in a manner similar to those in Lobell (2013), which assumed normally distributed measurement errors.

While our analysis focuses on temperature, the same approach can be applied to other environmental variables, such as precipitation, humidity, wind speed, or air quality, where interpolation, aggregation, and spatial smoothing are equally pervasive. The framework can also be used to assess the performance of alternative reanalysis data products or interpolation algorithms, to evaluate derived exposure measures such as heatwave duration or degree days, and to test the success of proposed bias-correction methods.

The breadth and scope of the measurement error issues we document in this paper preclude a single, easily implementable solution, but they also point to several promising directions for future work. First, improvements in monitoring and reporting systems could provide temperature data at finer spatial and temporal resolutions. Second, progress can build on recent advances in addressing measurement error in remote-sensing estimates (e.g., Alix-Garcia and Millimet (2023)), where discrepancies between proxies and ground truth are well recognized. Third, although we show in Appendix C.1 why standard instrumental variables approaches fail in this setting, variants that explicitly address the violated assumptions may still be viable. Fourth, and most directly aligned with our framework, is the correction-factor approach long used in industrial experiments. In Appendix C.2, we implement this approach using a multivariate adaptive regression spline model and obtain temperature slope estimates that are substantially less biased, highlighting a promising avenue for future research.

References

Alix-Garcia, J. and Millimet, D. L. (2023). Remotely incorrect? accounting for nonclassical measurement error in satellite data on deforestation. *J. Assoc. Environ. Resour. Econ.*, 10(5):1335–1367.

Andreou, E., Ghysels, E., and Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *J. Econom.*, 158(2):246–261.

Auffhammer, M. (2022). Climate adaptive response estimation: Short and long run impacts of climate change on residential electricity and natural gas consumption. *J. Environ. Econ. Manage.*, 114(102669):102669.

Auffhammer, M., Hsiang, S. M., Schlenker, W., and Sobel, A. (2013). Using weather data and climate model output in economic analyses of climate change. *Rev. Environ. Econ. Pol.*, 7(2):181–198.

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. *J. Polit. Econ.*, 124(1):105–159.

Barreca, A. and Schaller, J. (2020). The impact of high ambient temperatures on delivery timing and gestational lengths. *Nat. Clim. Chang.*, 10(1):77–82.

Baylis, P. (2020). Temperature and temperament: Evidence from twitter. *J. Pub. Econ.*, 184(104161):104161.

Berman, J. D., Bayham, J., and Burkhardt, J. (2020). Hot under the collar: A 14-year association between temperature and violent behavior across 436 US counties. *Environ. Res.*, 191.

Bernal, P., Mittag, N., and Qureshi, J. A. (2016). Estimating effects of school quality using multiple proxies. *Labour Econ.*, 39:1–10.

Black, D. A., Berger, M. C., and Scott, F. A. (2000). Bounding parameter estimates with nonclassical measurement error. *J. Am. Stat. Assoc.*, 95(451):739–748.

Black, D. A. and Smith, J. A. (2006). Estimating the returns to college quality with multiple proxies for quality. *J. Labor Econ.*, 24(3):701–728.

Borsky, S., Fesselmeyer, E., and Vogelsang, L. (2024). Urban heat and within-city residential sorting. *J. Env. Econ. Manage*, 127:103014.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Stat. Assoc.*, 90(430):443–450.

Burke, M., González, F., Baylis, P., Heft-Neal, S., Baysan, C., Basu, S., and Hsiang, S. (2018). Higher temperatures increase suicide rates in the united states and mexico. *Nat. Clim. Chang.*, 8(8):723–729.

Carleton, T., Jina, A., Delgado, M., Greenstone, M., Houser, T., Hsiang, S., Hultgren, A., Kopp, R. E., McCusker, K. E., Nath, I., et al. (2022). Valuing the global mortality consequences of climate change accounting for adaptation costs and benefits. *Q. J. Econ*, 137(4):2037–2105.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Chen, X., Hong, H., and Nekipelov, D. (2011). Nonlinear models of measurement errors. *J. Econ. Lit.*, 49(4):901–937.

Colmer, J. and Doleac, J. L. (2023). Access to guns in the heat of the moment: more restrictive gun laws mitigate the effect of temperature on violence. *Rev. Econ. Stats.*, pages 1–40.

Cook, N. and Heyes, A. (2020). Brain freeze: outdoor cold and indoor cognitive performance. *J. Env. Econ. Manage*, 101:102318.

Cramer, J. S. (1987). Mean and variance of R2 in small and moderate samples. *J. Econom.*, 35(2-3):253–266.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., and Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. *Int. J. Climatol.*, 28(15):2031–2064.

Deschênes, O. and Greenstone, M. (2007). The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather. *Am. Econ. Rev.*, 97(1):354–385.

Deschênes, O. and Greenstone, M. (2011). Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US. *Am. Econ. J. Appl. Econ.*, 3(4):152–185.

Dong, H. and Millimet, D. L. (2024). Embrace the noise: it is ok to ignore measurement error in a covariate, sometimes. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 188:608–627.

Durbin, J. and Watson, G. S. (1992). Testing for serial correlation in least squares regression. ii. In *Breakthroughs in Statistics: Methodology and Distribution*. Springer.

Fisher, A. C., Hanemann, W. M., Roberts, M. J., and Schlenker, W. (2012). The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather: Comment. *Am. Econ. Rev.*, 102(7):3749–3760.

Fowlie, M., Rubin, E., and Walker, R. (2019). Bringing satellite-based air quality estimates down to earth. *AEA Pap. Proc.*, 109:283–288.

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

Ghanem, D. and Smith, A. (2021). What are the benefits of high-frequency data for fixed effects panel models? *J. Assoc. Environ. Resour. Econ.*, 8(2):199–234.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *J. Polit. Econ.*, 127(4):1826–1863.

Graff Zivin, J., Hsiang, S. M., and Neidell, M. (2018). Temperature and human capital in the short and long run. *J. Assoc. Environ. Resour. Econ.*, 5(1):77–105.

Grainger, C. and Schreiber, A. (2019). Discrimination in ambient air pollution monitoring? *AEA Pap. Proc.*, 109:277–282.

Grieneisen, M. L. and Zhang, M. (2011). The current status of climate change research. *Nat. Clim. Chang.*, 1(2):72–73.

Griliches, Z. (1974). Errors in variables and other unobservables. *Econometrica*, 42(6):971.

Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *J. Econ. Perspect.*, 15(4):57–67.

Heilmann, K., Kahn, M. E., and Tang, C. K. (2021). The urban crime and heat gradient in high and low poverty areas. *J. Pub. Econ.*, 197:104408.

Hendricks, N. P. (2018). Potential benefits from innovations to reduce heat and water stress in agriculture. *J. Assoc. Environ. Resour. Econ.*, 5(3):545–576.

Heutel, G., Miller, N. H., and Molitor, D. (2021). Adaptation and the mortality effects of temperature across US climate regions. *Rev. Econ. Stat.*, 103:740–753.

Hsiang, S. (2016). Climate econometrics. *Annu. Rev. Resour. Economics*, 8(1):43–75.

Josephson, A., Michler, J. D., Kilic, T., and Murray, S. (2024). The mismeasure of weather: Using remotely sensed earth observation data in economic context. *arXiv:2409.07506*.

Khalaf, L., Kichian, M., Saunders, C. J., and Voia, M. (2021). Dynamic panels with MIDAS covariates: Nonlinearity, estimation and fit. *J. Econom.*, 220(2):589–605.

Klaiber, H. A., Abbott, J. K., and Smith, V. K. (2017). Some like it (less) hot: Extracting tradeoff measures for physically coupled amenities. *J. Assoc. Environ. Resour. Econ.*, 4(4):1053–1079.

Kolstad, C. D. and Moore, F. C. (2020). Estimating the economic impacts of climate change using weather observations. *Rev. Environ. Econ. Pol.*, 14(1):1–24.

Kydland, F. E. and Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica*, 50(6):1345–1370.

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

Lubotsky, D. and Wittenberg, M. (2006). Interpretation of regressions with multiple proxies. *Rev. Econ. Stat.*, 88(3):549–562.

Ma, Y. and Guttorp, P. (2013). Estimating daily mean temperature from synoptic climate observations. *Int. J. Climatol.*, 33(5):1264–1269.

Maccini, S. and Yang, D. (2009). Under the weather: Health, schooling, and economic consequences of early-life rainfall. *Am. Econ. Rev.*, 99(3):1006–1026.

Millimet, D. L. and Parmeter, C. F. (2022). Accounting for skewed or one-sided measurement error in the dependent variable. *Political Anal.*, 30(1):66–88.

Millimet, D. L. and Parmeter, C. F. (2025). The impact of measurement error on trends in earnings inequality in the usa. *Empir. Econ.*, pages 1–27.

Mira, A. (1999). Distribution-free test for symmetry based on bonferroni's measure. *J. Appl. Stat.*, 26(8):959–972.

Mullins, J. T. and White, C. (2019). Temperature and mental health: Evidence from the spectrum of mental health outcomes. *J. Health Econ.*, 68:102240.

Park, R. J., Goodman, J., Hurwitz, M., and Smith, J. (2020). Heat and learning. *Am. Econ. J.: Econ. Policy*, 12(2):306–39.

Pouliot, G. A. (2023). Spatial econometrics for misaligned data. *J. Econom.*, 232(1):168–190.

Proctor, J., Carleton, T., and Sum, S. (2023). Parameter recovery using remotely sensed variables. Technical report, National Bureau of Economic Research, Working Paper 30861.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *Am. Sociol. Rev*, 38(2):337–341.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annu. Rev. Econom.*, 8(1):341–377.

Schlenker, W. and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proc. Natl. Acad. Sci. U. S. A.*, 106(37):15594–15598.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stoklosa, J., Daly, C., Foster, S. D., Ashcroft, M. B., and Warton, D. I. (2015). A climate of uncertainty: accounting for error in climate variables for species distribution models. *Methods Ecol. Evol.*, 6(4):412–423.

Yang, Y., Jia, F., and Li, H. (2023). Estimation of panel data models with mixed sampling frequencies. *Oxf. Bull. Econ. Stat.*, 85(3):514–544.

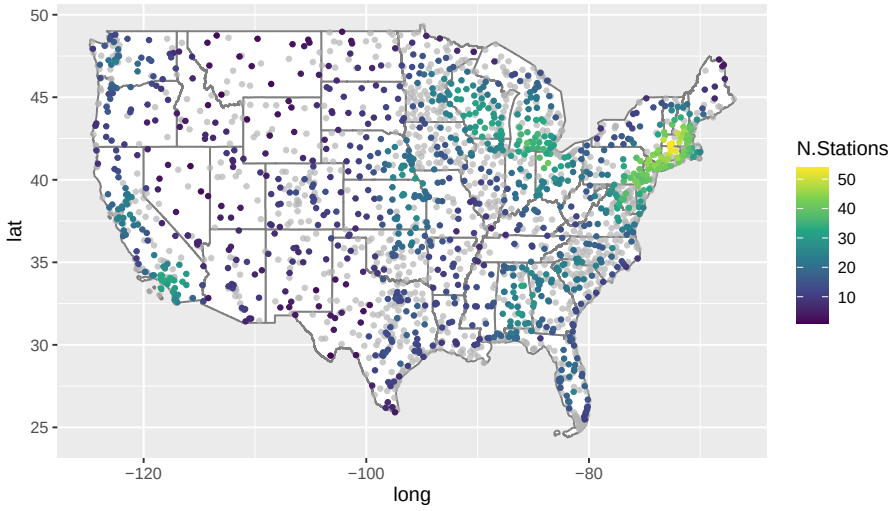
Appendices

A Details on Simulation Design

A.1 Sequential Error Introduction

Our panel dataset of 1,039 U.S. temperature monitors allows us to construct three sets of daily temperature proxies for each monitor denoted by T_{1d}, T_{2d} and T_{3d}. A sequential process introduces a different source of measurement error at each step. At each monitor, we estimate equation (4) three times, sequentially incorporating different sources of measurement error: (a) monitor interpolation, (b) temporal interpolation, and (c) binning.[30] For each step, we (i) generate the measurement error specific to that step; and (ii) analyze how this error affects the estimation of the response function given by equation (4), examining both the slope coefficient β_1 and the R-squared.

Figure A.1: Map of Sampled Monitors



Note: Gray dots indicate monitors excluded due to $> 5\%$ missing data or no adjacent monitors within 200km. Colored dots represent 1,039 monitors included in the analysis, with lighter colors indicating higher density of nearby monitors and darker colors showing areas of sparse coverage.

Adding monitor interpolation error. The first source of error comes from a weather monitor not being placed at the exact location where the outcome is observed. We start by using the closest monitor as the proxy and then compute interpolation errors at each of our ground truth monitors. For example, if the mean daily temperature at the San Francisco airport (SFO)

[30] While each subsequent proxy incorporates measurement errors from previous steps, creating a cumulative error structure, the magnitude of total error need not increase monotonically. This occurs because measurement errors from different sources can offset one another to different degrees.

is treated as the ground truth, then the nearest monitor to use for the proxy readings is 10 miles away at the San Carlos Airport (SQL):

$$T_{1d} = T_{mean,SQL,d}$$

We estimate equation (4) using this proxy to examine how this first source of error affects both the R-squared and β_1 estimates.

Adding temporal interpolation error. The second source of error stems from limitations in temporal resolution of temperature measurements. While our ground truth utilizes the average of 24-hourly readings, many analyses span longer periods when numerous stations only record daily maximum and minimum temperatures. Large-scale ensemble methods such as PRISM also typically employ this approach for estimating daily mean temperature. The standard calculation for mean temperature using this method is

$$T_{2d} = \frac{T_{max,SQL,d} + T_{min,SQL,d}}{2}$$

The cumulative measurement error at this stage combines both monitor and temporal interpolation errors. We re-estimate equation (4) using this new proxy to assess the combined impact of both error sources.

Incorporating binning error. The next issue considered arises from the common practice of binning temperature data that are themselves measured with error. Following standard convention, we define 10 binary indicators $\text{TBIN}_{j,d}$ for temperature ranges:

$$\text{TBIN}_{j,d} = \begin{cases} 1 & \text{if } T_{2d} \text{ falls in bin } j \\ 0 & \text{otherwise} \end{cases}$$

where bins are defined as: below 10°F, 10-20°F, ..., 80-90°F, and above 90°F. The corresponding regression model, , parallel to equation (4), is

$$y_d = \sum_{j=1}^{10} \alpha_j \text{TBIN}_{j,d} + \nu_d$$

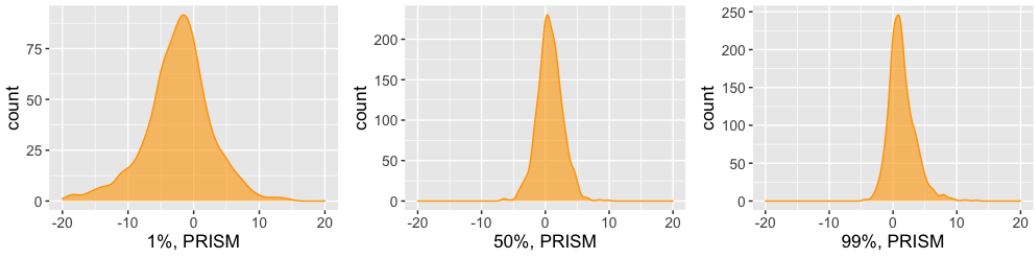
To keep this binned version comparable with our earlier continuous proxies, we also construct a third proxy T_{3d}, by assigning to each observation the mean temperature of the bin in which T_{2d} falls. This allows direct comparison of cumulative measurement error across T_{1d}, T_{2d} and T_{3d}. T_{3d} now reflects monitor interpolation, temporal interpolation, and binning error, which arises both from misclassification near bin boundaries and from the loss of within-bin variation. Binning implicitly assumes identical impacts for all temperatures within a bin, or equivalently,

linear responsiveness within each bin—a simplification that may not hold in practice.[31]

A.2 PRISM Proxy

For each weather station in our sample, we extract PRISM's estimated daily mean temperature for the station's grid cell as the proxy measure. PRISM's construction combines our first two sources of error: 1) "monitor" interpolation error from the smoothing algorithm, and 2) temporal interpolation error from obtaining an estimate of mean temperature by averaging daily maximum and minimum temperatures. This allows direct comparison between PRISM-based proxies and our previously constructed second stage proxies T_{2d}. The third stage proxy, T_{3d}, which examines how measurement error in T_{2d} translates into bin classification, is also directly comparable. However, a deeper methodological issue arises when comparing PRISM's performance to the closest monitor proxy. While the closest monitor proxy does not use information about temperature at the ground truth monitor's location, PRISM incorporates readings from all monitors in its construction—including the ground truth monitor itself. While PRISM's sophisticated spatial interpolation algorithms prevent exact reproduction of the ground truth measurements, inclusion of the ground truth monitor's temperature readings in its information set creates an inherent advantage for PRISM relative to the closest monitor approach.

Figure A.2: Measurement Errors at 1%, 50%, and 99% of Temperature Distribution for PRISM



B Measurement Error Properties

B.1 Measurement Error Varies with GIS Variables

To better understand the underlying drivers of these geographic patterns, we calculate the Root Mean Square Error (RMSE)[32] for each of our 1,039 weather stations by averaging squared measurement errors over all daily observations at that station. This station-level analysis allows us to systematically examine how measurement error relates to geographic features and monitoring network characteristics.

[31] This assumption can be justified only if the temperature–response relationship is homogeneous within each bin. For example, if temperatures are uniformly distributed and effects are linear within bins. More generally, binning smooths over nonlinearities or threshold effects. If responses change sharply around a specific cutoff (e.g., 88°F), a 80–90°F bin averages that effect across the interval, potentially masking it.

[32] While we also calculate standard deviations of the measurement errors, we focus on RMSE as it captures both the variance and bias in measurement errors while penalizing larger deviations more heavily.

Weather station density plays a crucial role in measurement errors of temperature, but affects different methods in distinct ways. Figure A.3 shows the relationship between measurement errors and weather station density across different distance bands (the total 0-200km range, 0-50km, 50-100km, 100-150km, and 150-200km). The top panel displays average RMSE as the Y axis and the bottom panel shows SD of measurement errors at each of 1,039 weather stations. The figure shows a consistent negative relationship between station counts and measurement errors across different distance bands. While the fitted lines appear visually similar, the error-reducing effect is strongest for stations within 50km. This is because the x-axis scales differ across panels - the 0-50km band spans only 10 stations, while outer bands span 15-20 stations. Thus, when normalized to the same scale, the 0-50km band shows the steepest negative slope, suggesting that having more stations in an immediate proximity matters most for accuracy. This relationship becomes progressively weaker at greater distances, as evidenced by the flatter slopes in the 100-150km and 150-200km bands, indicating diminishing returns from more distant stations.

The relationship between measurement error and distance to the nearest station (Figure A.4) further illustrates important differences between estimation methods. Traditional interpolation (top panel) shows a clear positive relationship - errors increase roughly linearly with distance, particularly beyond 100km. In contrast, PRISM (bottom panel) demonstrates remarkable stability, maintaining consistent error levels even at greater distances. This suggests PRISM's incorporation of additional geographic information helps mitigate distance-related degradation in accuracy. Traditional interpolation performs better at short distances (<50km) with a lower RMSE, while PRISM appears to offer more reliable estimates in sparsely monitored regions.

Elevation also strongly influences measurement errors (Figure A.5). Traditional interpolation shows a strong positive relationship with elevation, with errors increasing more rapidly at higher elevations and reaching 15-20°F in some cases. PRISM displays a more nuanced quadratic relationship - errors initially increase but then decrease at higher elevations, rarely exceeding 7°F. The methods' performance lines cross at around 1,600 feet elevation.

Examination of error distributions by land use patterns reveal distinct differences by interpolation methods (see Figure A.6). Traditional interpolation shows substantial variation in measurement errors (2.5-10°F), with natural landscapes like Evergreen Forest and Barren Land exhibiting the highest errors (8-10°F). In contrast, PRISM demonstrates more consistent but generally higher errors around 5°F across land types, suggesting its terrain-aware algorithm trades higher accuracy in specific landscapes for more uniform performance across terrain types.

To complement our graphical analysis, we conduct a regression analysis relating RMSE to geographic and network characteristics, with consistent results reported in Table A.1.[33] For each station, we regress measurement errors on a comprehensive set of variables, including elevation, distances to the coast, water, and roads, population density, land type indicators, and

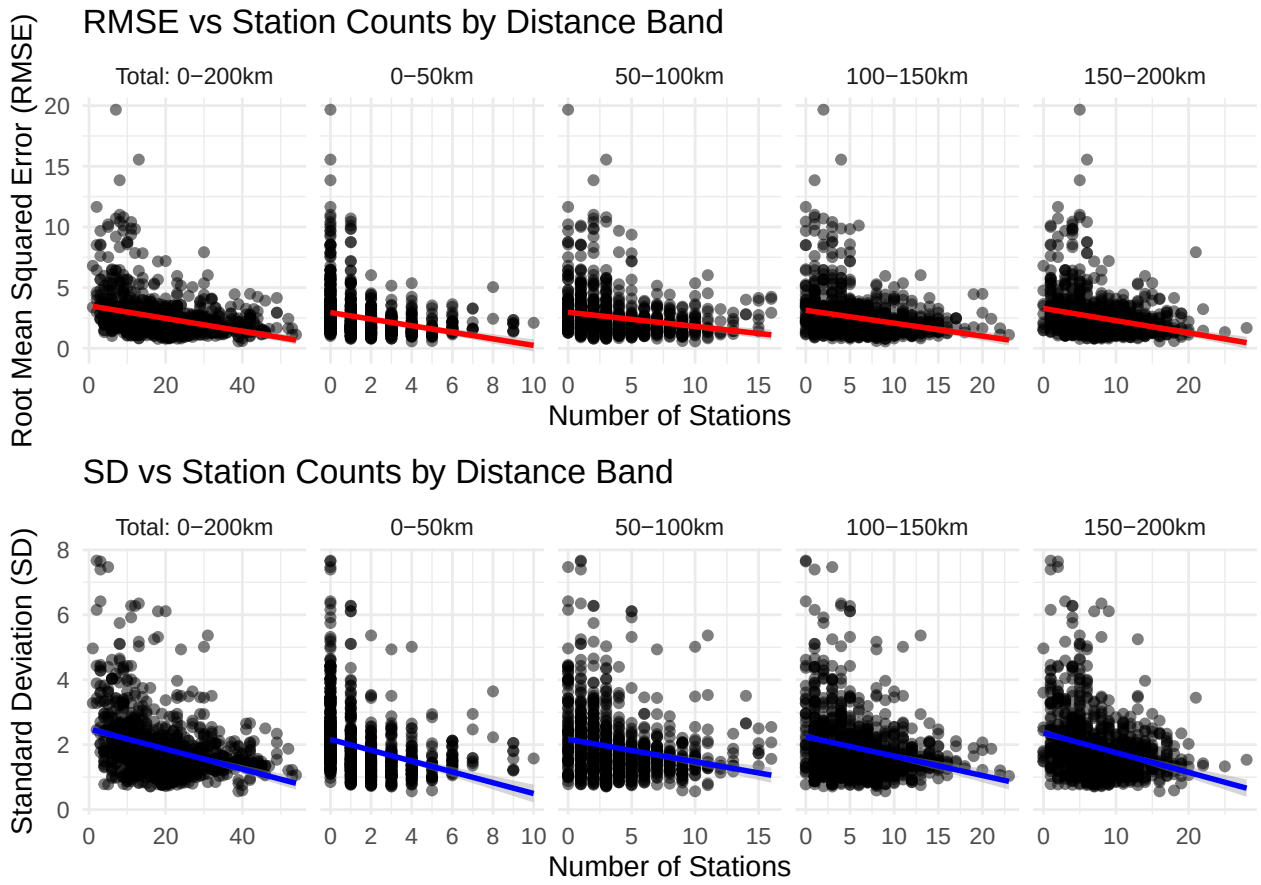
[33] A similar analysis using the standard deviation of station-level measurement errors is presented in Table A.2.

monitoring network features, such as the number of stations in different distance bands and the distance to the nearest station. Interpreting these regression results requires caution due to high correlations among geographic and network characteristics. For example, elevation, land use types, population density, and station density are all naturally interrelated - urban areas tend to be at lower elevations with denser monitoring networks. Nevertheless, the regression analysis helps confirm several key patterns. The regression analysis confirms that elevation strongly influences traditional interpolation errors but has minimal effect on PRISM errors, while distance and station density variables exhibit the patterns observed in our earlier analysis.

Interestingly, geographical features explain variations in PRISM measurement errors better than in traditional interpolation method (adjusted R-squared 0.532 versus 0.418). This higher explanatory power, while counterintuitive given PRISM's geographic corrections, suggests potential systematic over- or under-adjustment for certain terrain characteristics. The geographic variables may capture aspects of terrain complexity not fully incorporated into PRISM's algorithm, warranting further investigation of these interactions.

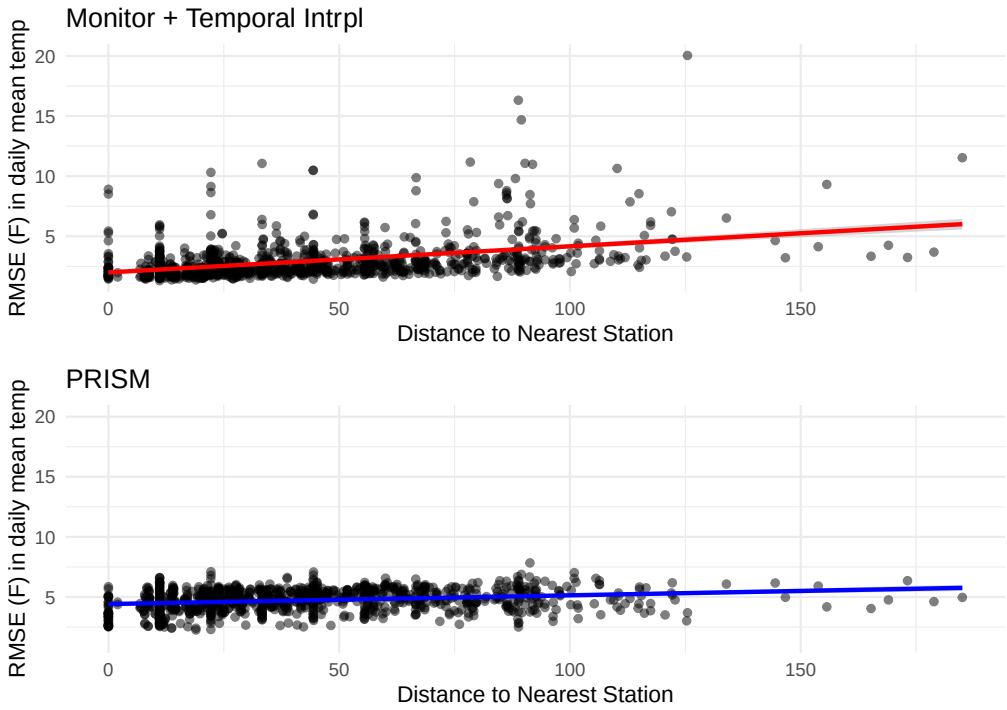
These findings highlight that measurement errors have strong systematic components related to geography and monitoring network characteristics. The choice between estimation methods should consider local conditions: traditional interpolation might be preferable for low-elevation, densely monitored urban areas, while PRISM offers more reliable estimates in regions with complex terrain or sparse monitoring networks.

Figure A.3: RMSE and Standard Deviation vs Station Counts by Distance Band



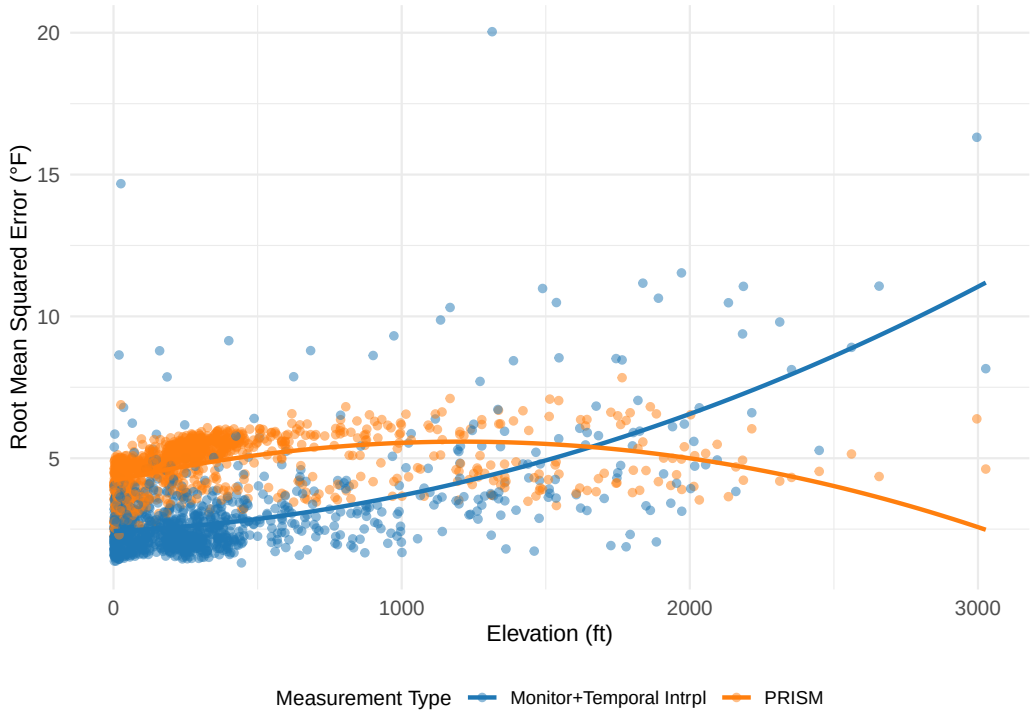
Note: Relationship between measurement errors and weather station density across different distance bands (the total 0-200km range, 0-50km, 50-100km, 100-150km, and 150-200km). The measurement errors shown are generated from traditional interpolation (Monitor + Temporal Interpolation). Top panel displays average RMSE and the bottom panel shows SD of measurement errors at each of 1,039 weather stations. For each distance band, the plots show how measurement errors vary with the number of available weather stations. Red and blue lines represent linear fits for RMSE and SD, respectively. The consistent negative slopes across all distance bands indicate that having more nearby stations generally improves temperature prediction accuracy.

Figure A.4: RMSE vs Distance to Nearest Weather Station



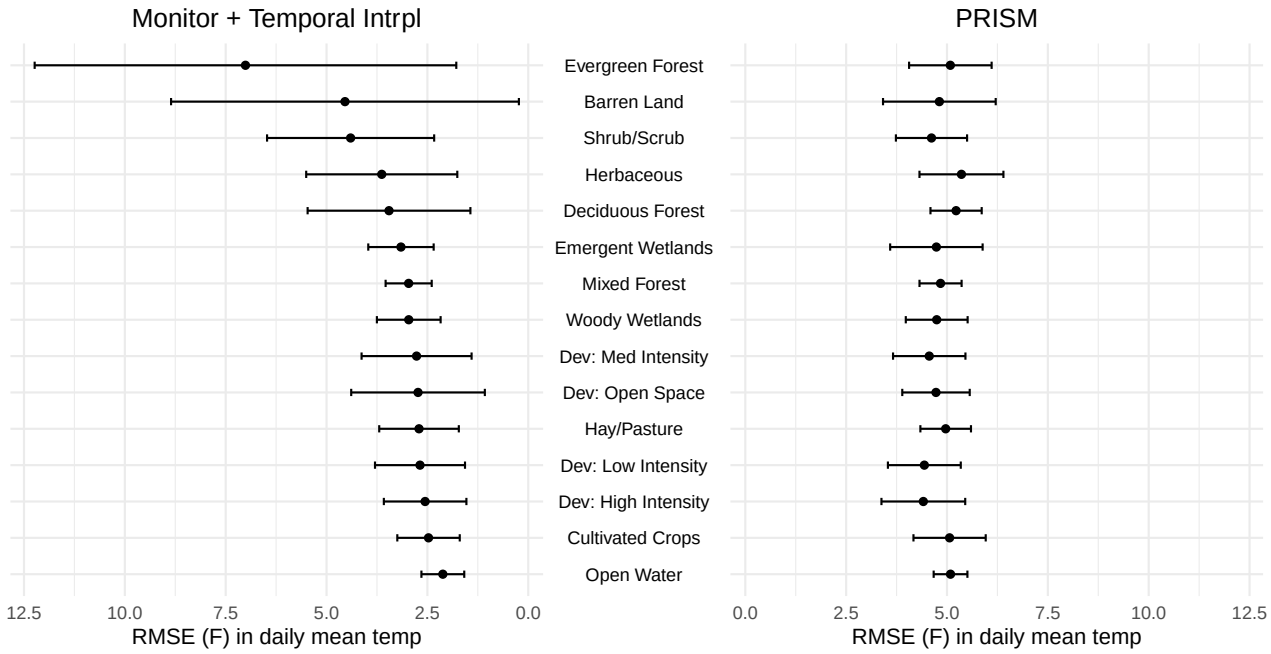
Note: Root Mean Square Error (RMSE) in daily mean temperature estimates as a function of distance to the nearest weather station. The visualization consists of two panels: the top panel shows errors from traditional interpolation (Monitor + Temporal Interpolation, red trend line), while the bottom panel shows PRISM estimation errors (blue trend line). Each point represents a weather station, with its position indicating both the distance to its nearest neighboring station (x-axis) and the associated measurement error (y-axis).

Figure A.5: RMSE vs. Elevation



Note: Root Mean Square Error (RMSE) in daily mean temperature as a function of elevation by proxy types: traditional interpolation (Monitor + Temporal Interpolation, shown in blue) and PRISM (shown in orange). The plot includes quadratic fit lines to capture potential nonlinear relationships between elevation and measurement error. Each point represents a weather station, positioned according to its elevation (x-axis) and associated measurement error (y-axis).

Figure A.6: RMSE by Land Use Type



Note: Measurement errors (RMSE) in daily mean temperature's variability across different land use categories, with traditional interpolation method (Monitor + Temporal Interpolation) in the left panel and with PRISM estimates in the right panel. Land use categories are ordered by increasing RMSE in the interpolation method, ranging from developed urban areas at the bottom to more complex terrains like barren land at the top. Each point represents the mean RMSE for that land use type, with horizontal error bars showing one standard deviation of uncertainty.

B.2 Details on Measurement Error Tests

B.2.1 Single Monitor Tests of the Classical Measurement Error Assumptions

Tests on our complete dataset of over 1,000 weather stations overwhelmingly reject the classical assumptions at any conventional significance level, in large part due to the power our sample size of several million observations provides. As studies are sometimes focused on a single location and a relatively short time series (e.g., Borsky et al. (2024); Cook and Heyes (2020)), we focus our analysis on four representative U.S. weather stations that capture diverse geographical and climatic conditions: San Francisco, CA; Boulder, CO; Chicago, IL; and Palm Beach, FL.

To further diminish power, instead of using all available daily observations ($\approx$ 4,017 per station), we repeatedly sample 500 consecutive days from each station and average the test results over 100 iterations. This sampling approach preserves the temporal dependency structure necessary for serial correlation analysis while providing more reasonable statistical power for hypothesis testing. We examine mean zero, serial correlation, independence from latent

Table A.1: Effect of GIS variables on RMSE

	Dependent variable:						
	Monitor Intrpl	+Temporal Intrpl	PRISM	0-50km	50-100km	100-150km	150-200km
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Elevation	0.002***	0.002***	−0.00001	0.002***	0.003***	0.002***	0.003***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)
Dist to Coast	−0.001***	−0.001***	0.001***	−0.001***	−0.001***	−0.001***	−0.001***
	(0.0001)	(0.0001)	(0.00005)	(0.0001)	(0.0001)	(0.0002)	(0.0002)
Dist to Water	−0.012*	−0.011*	0.010***	−0.013*	−0.020**	−0.022*	−0.027**
	(0.007)	(0.007)	(0.004)	(0.007)	(0.009)	(0.012)	(0.013)
Dist to Roads	−0.019	−0.011	−0.001	0.020	−0.019	−0.007	−0.005
	(0.021)	(0.020)	(0.010)	(0.024)	(0.030)	(0.033)	(0.036)
Population Density	−0.00002	−0.00003	−0.00002	−0.00002	0.0001	−0.00002	−0.00001
	(0.00004)	(0.00004)	(0.00002)	(0.00003)	(0.00004)	(0.0001)	(0.0001)
Stations 0-50 km	0.030	0.028	−0.016	0.003	−0.009	0.104*	0.016
	(0.033)	(0.031)	(0.016)	(0.029)	(0.039)	(0.054)	(0.057)
Stations 50-100 km	0.027	0.026	0.004	0.033**	0.010	0.063**	0.095***
	(0.018)	(0.017)	(0.009)	(0.016)	(0.021)	(0.028)	(0.031)
Stations 100-150 km	−0.013	−0.011	0.022***	−0.007	0.003	−0.039*	−0.051**
	(0.014)	(0.013)	(0.007)	(0.012)	(0.016)	(0.022)	(0.024)
Stations 150-200 km	0.001	0.002	0.018***	−0.002	0.002	−0.032*	−0.017
	(0.012)	(0.011)	(0.006)	(0.011)	(0.014)	(0.019)	(0.021)
Dist to Nearest Station	0.017***	0.015***	0.002**	0.013***	−0.003	0.002	0.0003
	(0.002)	(0.002)	(0.001)	(0.003)	(0.002)	(0.003)	(0.003)
Land Type:							
Dev: Open Space	0.643	0.479	0.014	0.509	0.283	0.661	0.859
	(0.593)	(0.568)	(0.293)	(0.600)	(0.685)	(0.950)	(1.037)
Dev: Low Intensity	0.683	0.508	−0.140	0.664	0.433	0.947	1.187
	(0.597)	(0.573)	(0.295)	(0.604)	(0.690)	(0.957)	(1.044)
Dev: Med Intensity	0.798	0.619	−0.042	0.757	0.551	0.783	1.485
	(0.599)	(0.574)	(0.296)	(0.604)	(0.693)	(0.960)	(1.047)
Dev: High Intensity	0.628	0.471	−0.106	0.530	0.359	0.902	1.282
	(0.605)	(0.580)	(0.299)	(0.610)	(0.700)	(0.970)	(1.058)
Barren Land	1.979***	1.814**	0.399	0.731	2.340***	2.654**	4.206***
	(0.752)	(0.721)	(0.371)	(0.736)	(0.869)	(1.237)	(1.315)
Deciduous Forest	1.169*	0.940	0.095	1.494**	0.511	0.581	0.438
	(0.702)	(0.673)	(0.347)	(0.705)	(0.811)	(1.126)	(1.228)
Evergreen Forest	2.963***	2.881***	0.528	3.283***	2.480***	1.776	0.693
	(0.727)	(0.698)	(0.359)	(0.792)	(0.841)	(1.187)	(1.272)
Mixed Forest	0.979	0.933	0.137	1.007	0.926	1.629	1.182
	(0.882)	(0.846)	(0.436)	(0.842)	(1.019)	(1.413)	(1.542)
Shrub/Scrub	0.239	0.162	−0.166	0.619	−0.327	1.020	1.108
	(0.622)	(0.596)	(0.307)	(0.645)	(0.729)	(1.006)	(1.091)
Herbaceous	0.757	0.585	0.287	0.802	0.469	0.306	0.965
	(0.606)	(0.582)	(0.300)	(0.615)	(0.704)	(0.972)	(1.062)
Hay/Pasture	0.756	0.669	0.104	0.563	0.432	0.472	0.659
	(0.605)	(0.580)	(0.299)	(0.613)	(0.700)	(0.970)	(1.058)
Cultivated Crops	0.609	0.446	−0.175	0.512	0.373	0.844	1.297
	(0.612)	(0.587)	(0.303)	(0.620)	(0.708)	(0.982)	(1.072)
Woody Wetlands	1.177	0.990	−0.158	0.877	0.877	2.104	1.899
	(0.812)	(0.779)	(0.401)	(0.940)	(0.961)	(1.302)	(1.421)
Emergent Wetlands	1.350*	1.230*	−0.225	1.312*	1.869**	2.379**	1.353
	(0.750)	(0.720)	(0.371)	(0.761)	(0.893)	(1.202)	(1.349)
Constant	0.734	1.347**	3.653***	0.907	2.359***	3.019***	3.409***
	(0.614)	(0.589)	(0.304)	(0.623)	(0.715)	(0.988)	(1.075)
Observations	1,039	1,039	1,039	688	959	997	1,016
Adjusted R^2	0.430	0.418	0.532	0.308	0.339	0.212	0.278

Note: Cross-sectional regressions at weather station level (N=1,039). Dependent variables are station-specific standard deviations of temperature measurement errors averaged over 2010-2020. For each station, standard deviations are calculated using different temperature measurement approaches: Monitor interpolation, + Temporal Interpolation, PRISM, and measurements within different distance bands (0-50km, 50-100km, 100-150km, 150-200km). Geographic variables include elevation, distances to coast/water/roads, population density, forest coverage, monitoring network characteristics (number of stations in different distance bands and distance to nearest station), and land type indicators.

Table A.2: Effect of GIS variables on Standard Deviation of Measurement Errors

	Dependent variable:						
	Closest	MinMax	PRISM	0-50km	50-100km	100-150km	150-200km
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Elevation	0.001***	0.001***	−0.00001	0.001***	0.001***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Dist to Coast	−0.0004***	−0.0003***	0.001***	−0.0005***	−0.001***	−0.001***	−0.001***
	(0.0001)	(0.0001)	(0.00005)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Dist to Water	−0.006	−0.005	0.010***	−0.009**	−0.008*	−0.011	−0.023***
	(0.004)	(0.004)	(0.003)	(0.004)	(0.005)	(0.007)	(0.008)
Dist to Roads	−0.014	−0.015	0.003	0.018	−0.019	0.0004	−0.017
	(0.012)	(0.010)	(0.010)	(0.016)	(0.016)	(0.020)	(0.022)
Population Density	−0.00001	−0.00002	−0.00002	−0.00001	0.00001	−0.00003	−0.00001
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00004)	(0.00004)
Stations 0-50 km	0.018	0.011	−0.015	0.008	0.009	0.076**	0.034
	(0.019)	(0.017)	(0.016)	(0.019)	(0.021)	(0.032)	(0.035)
Stations 50-100 km	0.016	0.010	0.005	0.019*	0.024**	0.088***	0.082***
	(0.010)	(0.009)	(0.009)	(0.010)	(0.011)	(0.017)	(0.019)
Stations 100-150 km	−0.006	−0.002	0.022***	−0.003	−0.003	−0.033**	−0.018
	(0.008)	(0.007)	(0.007)	(0.008)	(0.009)	(0.013)	(0.015)
Stations 150-200 km	−0.009	−0.007	0.019***	−0.004	−0.005	−0.026**	−0.023*
	(0.007)	(0.006)	(0.006)	(0.007)	(0.007)	(0.011)	(0.013)
Dist to Nearest Station	0.013***	0.011***	0.002***	0.008***	0.001	0.003	0.002
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)
Land Type:							
Dev: Open Space	0.205	0.133	−0.00001	0.069	0.232	0.508	0.310
	(0.341)	(0.300)	(0.292)	(0.391)	(0.368)	(0.563)	(0.631)
Dev: Low Intensity	0.333	0.215	−0.154	0.220	0.379	0.740	0.474
	(0.344)	(0.302)	(0.294)	(0.394)	(0.371)	(0.567)	(0.636)
Dev: Med Intensity	0.387	0.275	−0.065	0.250	0.436	0.804	0.774
	(0.345)	(0.303)	(0.295)	(0.393)	(0.372)	(0.569)	(0.638)
Dev: High Intensity	0.284	0.202	−0.117	0.167	0.387	0.817	0.720
	(0.348)	(0.306)	(0.298)	(0.397)	(0.376)	(0.574)	(0.644)
Barren Land	0.351	0.155	−0.019	0.026	0.393	0.949	0.794
	(0.433)	(0.380)	(0.370)	(0.480)	(0.467)	(0.733)	(0.801)
Deciduous Forest	0.652	0.432	0.002	0.851*	0.525	0.406	0.320
	(0.404)	(0.355)	(0.346)	(0.459)	(0.436)	(0.667)	(0.748)
Evergreen Forest	0.821*	0.703*	0.252	1.425***	0.760*	0.961	0.281
	(0.419)	(0.368)	(0.358)	(0.516)	(0.452)	(0.703)	(0.775)
Mixed Forest	0.107	0.005	0.123	0.035	0.181	0.478	0.074
	(0.508)	(0.446)	(0.434)	(0.548)	(0.547)	(0.837)	(0.939)
Shrub/Scrub	0.376	0.357	−0.234	0.342	0.421	1.040*	1.088
	(0.358)	(0.314)	(0.306)	(0.420)	(0.392)	(0.596)	(0.664)
Herbaceous	0.563	0.430	0.253	0.479	0.630*	0.662	0.719
	(0.349)	(0.306)	(0.298)	(0.401)	(0.378)	(0.576)	(0.647)
Hay/Pasture	0.264	0.166	0.079	0.033	0.256	0.450	0.225
	(0.348)	(0.306)	(0.298)	(0.399)	(0.376)	(0.574)	(0.645)
Cultivated Crops	0.219	0.110	−0.203	0.055	0.271	0.634	0.853
	(0.353)	(0.309)	(0.301)	(0.404)	(0.380)	(0.582)	(0.653)
Woody Wetlands	0.742	0.641	−0.198	0.295	0.822	1.658**	1.274
	(0.468)	(0.411)	(0.400)	(0.612)	(0.516)	(0.771)	(0.865)
Emergent Wetlands	0.858**	0.720*	−0.312	0.539	1.495***	2.468***	0.997
	(0.432)	(0.379)	(0.369)	(0.496)	(0.480)	(0.712)	(0.822)
Constant	1.012***	1.580***	3.613***	1.173***	1.731***	2.093***	2.944***
	(0.354)	(0.310)	(0.302)	(0.406)	(0.384)	(0.585)	(0.655)
Observations	1,039	1,039	1,039	688	959	997	1,016
Adjusted R^2	0.407	0.412	0.543	0.217	0.265	0.188	0.177

Note: Dependent variables are standard deviations of measurement errors calculated using different temperature measurement approaches: Monitor interpolation, + Temporal Interpolation, PRISM, and measurements within different distance bands (0-50km, 50-100km, 100-150km, 150-200km). Geographic variables include elevation , distances to coast/water/roads, population density, forest coverage, monitoring network characteristics (number of stations in different distance bands and distance to nearest station), and land type indicators. Robust standard errors shown in parentheses.

variable, heteroskedasticity, symmetry and normality assumptions. We test these properties in five different contexts: monitor interpolation, temporal interpolation (adding daily min-max average), binning procedure, PRISM estimation, and PRISM with binning. The averaged test statistics are reported in Table 3.

1. $E(u_{id}) = 0$**.** We test the mean zero assumption with t-tests, with results shown in Table 3 columns (1)-(3). Traditional interpolation approaches exhibit substantial variation in mean errors across the four sampled locations, revealing important geographical patterns. San Francisco shows small errors with monitor interpolation (0.04°F), which is a result of positive and negative errors cancel out across different days, and larger errors when temporal interpolation is added (0.81°F), suggesting that the simple averaging of daily minimums and maximums introduces systematic bias in such coastal areas. Boulder presents the most extreme case, with monitor interpolation methods substantially overestimating temperature by approximately 15°F, likely due to its complex mountainous terrain and elevation differences between stations. Chicago and Palm Beach show more modest but still significant errors, with monitor interpolation underestimating temperatures by 1.36°F and 2.56°F respectively. PRISM generally produces smaller mean errors across locations, with this improvement particularly dramatic in Boulder where the error reduces to about 1°F. The p-values from t-tests suggest that the mean zero assumption is frequently rejected, though some methods (particularly PRISM in Chicago and Palm Beach) show higher p-values (>0.10), indicating potentially unbiased measurement in these areas.

Remember from equation (3), it is the variance of measurement error, rather than its mean, that plays the crucial role in determining estimation bias in regression analyses. The variance term expressed in standard deviation form is reported in column (2) and reveal distinct patterns across methods and locations. Traditional monitor interpolation shows location-dependent precision, with standard deviations ranging from very precise in Chicago (0.90°F) to much larger in Boulder (5.02°F). Adding temporal interpolation consistently increases the standard deviation across all locations - for instance, from 1.68°F to 2.16°F in San Francisco and from 0.90°F to 1.56°F in Chicago. PRISM, as a spatial smoothing estimator, shows larger but more uniform standard deviations across locations (2.60-6.28°F). The binning procedure further increases error variance in all cases, with standard deviations growing by roughly 50% when applied to either traditional interpolation or PRISM estimates.

2. Serial Correlation. To test the independence assumption of classical measurement error, we need to explicitly define and test against potential dependency structures. Since temperature measurement errors are generated over time, serial correlation is a natural alternative dependency structure to examine. We employ the Durbin-Watson test (Durbin and Watson, 1992) and to maintain the temporal structure necessary for this analysis, we use 500 consecutive daily observations in each sample and estimate a simple autoregressive model where the current period's error is regressed on its one-day lag.

The results, shown in Table 3 columns (4)-(5), reveal strong serial dependence in measurement errors, particularly for simpler interpolation methods. San Francisco shows the strongest serial correlation (0.76 for monitor interpolation), likely due to persistent marine layer effects that create systematic differences between coastal and inland stations. Chicago and Palm Beach exhibit moderate but still significant correlation (0.44 and 0.51 respectively), suggesting that measurement errors tend to persist over consecutive days even in relatively flat terrain. Boulder shows strong serial correlation (0.62) with traditional interpolation method, indicating that topography-induced measurement errors are also highly persistent. PRISM substantially reduces serial correlation across all locations, with coefficients near zero and often insignificant, particularly evident in Chicago and Palm Beach. This improvement suggests that PRISM's incorporation of elevation and other geographic factors helps break down the systematic patterns in measurement errors. The binning procedure consistently reduces serial correlation across all locations. This is not surprising, as binning always involves some loss of information in the original temperature variable through its discretization.

3. $\mathbf{cov}(u, T^*) = 0$. As shown in equation (2), when measurement errors are correlated with true temperatures, the classical measurement error result of attenuation bias may no longer hold. When $\text{cov}(u, T^*)$ is negative and large in magnitude, the estimated coefficient can be larger than the true parameter, leading to an overestimate rather than the downward bias associated with attenuation. The results in columns (6)-(7) reveal that this correlation is generally statistically distinct from zero, indicating violations of this classical assumption, although there is notable variation across locations and the particular method.

Boulder shows the strongest positive correlations (up to 0.46 with temporal interpolation), indicating that measurement errors are likely to systematically vary with temperature levels in complex terrain. Temperature differences between stations also become more pronounced at certain temperature ranges, possibly due to elevation-dependent temperature gradients. San Francisco shows moderate positive correlations with traditional interpolation method (0.15-0.28), while Chicago and Palm Beach tend to show weaker and often negative correlations, which are not statistically significant at the p=0.10 level.

Measurement errors associated with PRISM proxy tends to exhibit negative correlations with true temperature.

4. Heteroskedasticity. We test for heteroskedasticity using the Breusch-Pagan test with month indicators as regressors,[34] as reported in column (8). The test strongly rejects the null hypothesis of homoskedastic errors across all sampled locations and methods, indicating that error variances systematically vary over time.

[34]We use month indicators here, but error variances could potentially be functions of other temporal and geographic variables. Our choice of month indicators provides a straightforward test of time-varying error variance, though other specifications could reveal different patterns of heteroskedasticity.

This pattern is particularly pronounced in Boulder (p-values ≈ 0) but less severe in Chicago, especially with simpler interpolation methods (p-values > 0.20). San Francisco shows strong heteroskedasticity with traditional interpolation method (p-values < 0.03) but somewhat reduced with PRISM (p-value $= 0.06$), suggesting that PRISM's approach may help stabilize error variance in coastal areas. Palm Beach shows varying degrees of heteroskedasticity, with temporal interpolation and binning showing less severe patterns (p-values > 0.09).

These findings indicate that the classical assumption of constant error variance is violated in temperature measurement, with the severity of the violation varying by both location and measurement method. The presence of heteroskedasticity, across locations and seasons, suggests that measurement error precision varies systematically, and hence systematically influences the reliability of temperature impact estimates over space and time.

5. Symmetry. While neither symmetry nor its stronger counterpart, normality, is strictly necessary for classical measurement error assumptions, these properties are often assumed in practice, making formal testing valuable. Figure 3 illustrates measurement errors calculated from 2010-2020 data for the four representative cities. Rows (A)-(C) show how measurement errors accumulate as additional sources are included. Table 3 reports skewness and symmetry tests (columns 9–10) of Mira (1999).

The skewness tests in column (9) frequently rejected the null of a zero third moment, while the symmetry test, which evaluates the entire distribution rather than a single moment, rarely rejects symmetry. This suggests that while small mean-level imbalances exist, overall error distributions remain approximately mirror-symmetric around their centers. Boulder provides a clear illustration: its nearest-monitor proxy shows pronounced asymmetry due to complex topography, but PRISM substantially corrects this, yielding near-symmetric errors. Shapiro-Wilk tests (column 11) of normality, by contrast, reject the appropriateness of normality assumptions at the p<.01 level for most situations considered.

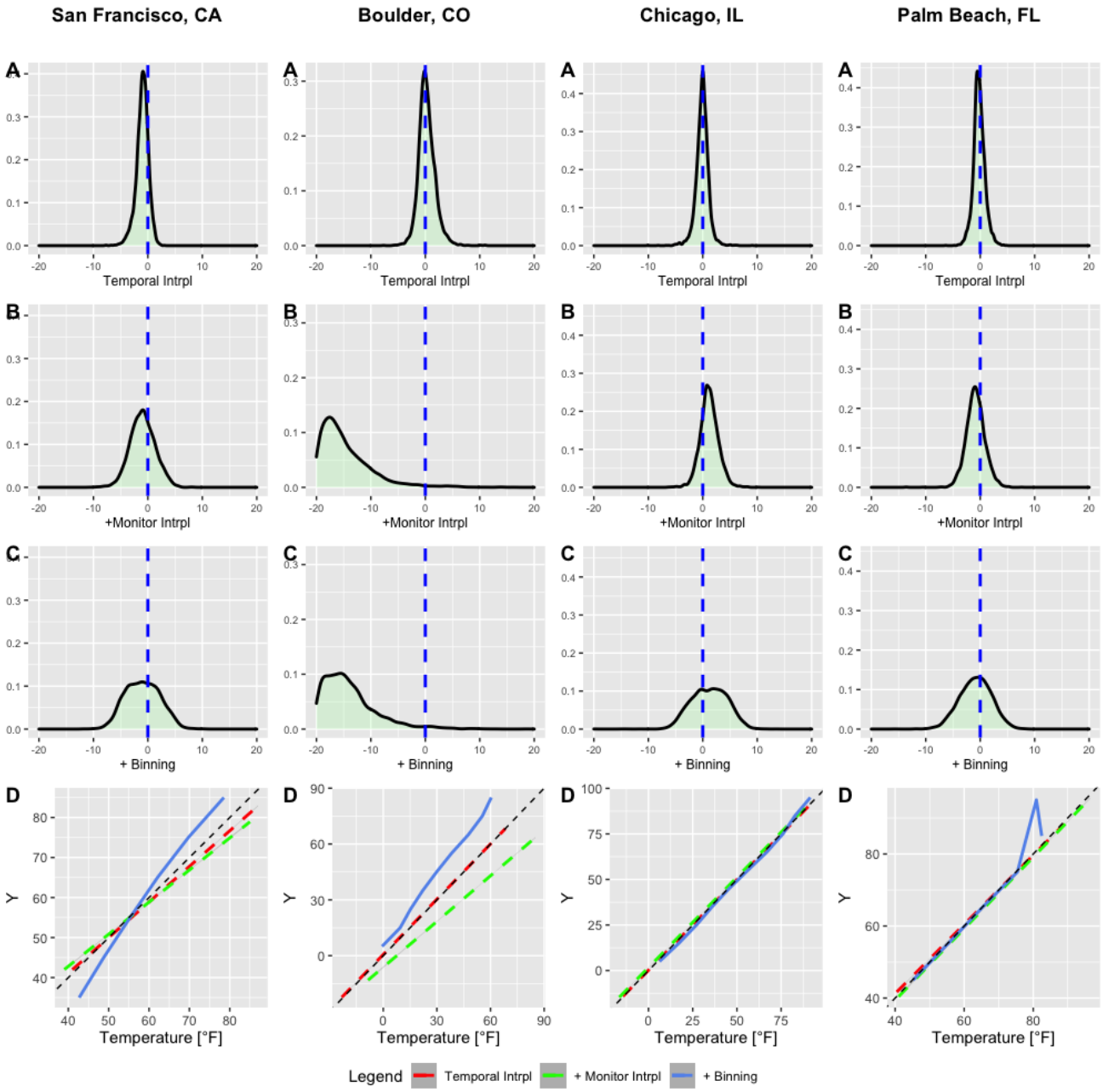
Taken together, these results demonstrate that temperature measurement errors generally violate the classical measurement error assumptions even in moderately-sized samples at individual monitors. The strength and importance of these violations vary systematically across locations and measurement methods, with some methods performing better (producing less biased estimates) than others depending on the specific assumption and location characteristics.

C Discussion on Correction Methods

C.1 The Instrumental Variable Approach

The instrumental variable (IV) approach is commonly used to correct for measurement error bias and as such it deserves further discussion in the specific context of temperature when there are

Figure A.7: Measurement Errors' Impact at Four Representative Cities – Alternative Order



multiple measures of temperature at a given location i on day d, one measurement Z_{id} can serve as the instrument for another measurement T_{id}.[35] The first stage regression in the IV approach is as follows:

$$T_{id} = \gamma_0 + \gamma_1 Z_{id} + \nu_{id}$$

While the IV approach is often effective in addressing classical measurement error, it becomes more problematic when the measurement error is non-classical. Three conditions must be met for the IV approach to be consistent:

1. Z_{id} is independent of the idiosyncratic error in the temperature response function.
2. Z_{id} is independent of the measurement error in T_{id}.
3. Z_{id} is related to T_{id}.

$$\hat{\beta}_1^{IV} \to \frac{\widehat{\text{cov}}(Y, Z)}{\widehat{\text{cov}}(T, Z)} = \beta_1 \times \frac{\widehat{\text{cov}}(T^*, Z)}{\widehat{\text{cov}}(T^*, Z) + \widehat{\text{cov}}(u, Z)} \tag{A.1}$$

The second condition is unlikely to hold in the context of estimating climate or temperature impact. As shown in sections 3, empirically constructed measurement errors in temperature observations are correlated with the true temperature and exhibit geographical and seasonal patterns. According to equation (A.1), the IV estimator converges to the true parameter multiplied by a ratio. When the measurement error in T_{id} is correlated with the instrument Z_{id} (i.e., $\widehat{\text{cov}}(u, Z) \neq 0$), the IV approach is no longer consistent. Furthermore, the size and sign of the measurement error bias depend on the relative magnitude of $\widehat{\text{cov}}(u, Z)$.

Although it is theoretically possible to construct a Z_{id} that satisfies all three conditions, do so is likely to be challenging in practice. Alternatively, one could obtain consistent estimates for $\widehat{\text{cov}}(T^*, Z)$ and $\widehat{\text{cov}}(u, Z)$ to recover the true β_1, a task we leave for future research.

C.2 Correction Factor Method for Non-Classical Measurement Error

This section describes the correction factor method used to adjust for non-classical measurement error and provides the simulation example corresponding to the main paper. The starting point is the probability limit expression when measurement error is non-classical:

$$\hat{\beta}_{\text{OLS}} \xrightarrow{p} \beta \frac{\text{var}(T^*) + \text{cov}(T^*, u)}{\text{var}(T^*) + \text{var}(u) + 2\,\text{cov}(T^*, u)} = \beta(1 - CF) \tag{A.2}$$

where

$$CF \equiv \frac{\text{cov}(u, T)}{\text{var}(T)}$$

[35] We assume the same measurement error structure as before: $T_{id} = T_{id}^* + u_{id}$

is the coefficient from regressing measurement error u on the proxy T. This expression makes clear that, if a consistent sample analogue of CF can be constructed, then the true parameter can be recovered through $\beta = \hat{\beta}_{\text{OLS}}/(1 - CF)$. The correction factor method therefore proceeds by first estimating CF where measurement error is observable and then predicting it for arbitrary locations and dates.

To illustrate this approach, we implement a simulation exercise parallel to the main paper. We designate 48 monitors as the testing sample and use the remaining monitors in the main analysis as the training sample. For each testing monitor i and month m, we observe true temperature T^*_{id} from daily station and construct the proxy temperature T_{id} using the nearest neighbor interpolation method. The simulated outcome is generated from the true data-generating process:

$$Y_{id} = T^*_{id}$$

so that the true slope equals one. Estimating the misspecified regression:

$$Y_{id} = \beta_0 + \beta_{im} T_{id} + e_{id}$$

produces an OLS estimate $\tilde{\beta}_{\text{OLS},im}$, whose bias is driven entirely by the measurement error in the proxy.

Having defined the testing environment, the next step is to construct the correction factor using the training monitors, where measurement error can be directly observed. Daily measurement error is computed as $u_{id} = T_{id} - T^*_{id}$. To estimate a stable correction factor, we pool all daily observations within the station month cell (i, m), yielding roughly 30 days per month over 11 years—about 330 observations per regression. For each station i and month m, we estimate

$$u_{id} = CF_{im} \times T_{id} + \varepsilon_{id}$$

with a separate regression run for every (i,m). The resulting station-month specific coefficients $\widetilde{CF}_{im}$ provide empirical estimates of how measurement error covaries with the proxy.

The final step is to extrapolate these estimated correction factors to unobserved locations. Because CF_{im} exhibits substantial geographic and seasonal heterogeneity, we model it as a function of observable characteristics:

$$CF_{im} = f(\text{GIS}_i,\ \text{time}_m)$$

using multivariate adaptive regression splines. The predictors include distance to the coast, elevation, distance to water and primary roads, population density, land-cover indicators, and month-of-year effects. The resulting predictions are then evaluated at the 48 testing monitors, yielding correction factors CF_{im} for each station–month in the test set.

Finally, for each test monitor and month, we compute the corrected estimate:

$$\tilde{\beta}_{\text{corrected},im} = \frac{\tilde{\beta}_{\text{OLS},im}}{1 - CF_{im}}.$$

Figure A.8 compares how the correction factor affects the estimated slope coefficients. The uncorrected nearest-neighbor estimates exhibit substantial attenuation, with the distribution centered well below the true value of one. After applying the correction factor, the distribution shifts markedly toward the true parameter, with the mean of the uncorrected parameter estimates being 0.928 and that of our corrected version being 0.994.

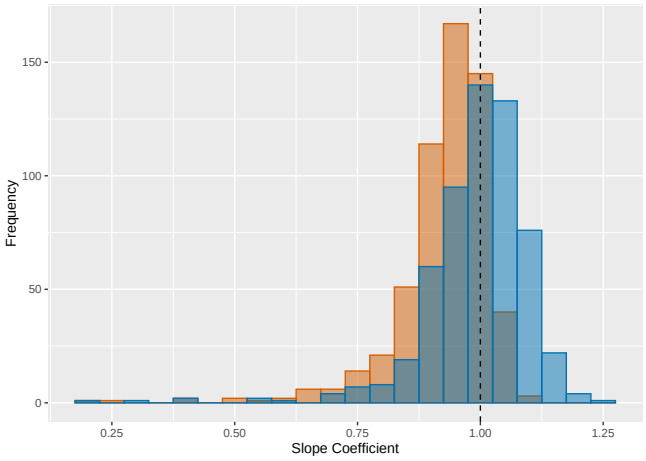


Figure A.8: Distribution of Estimated Slope Coefficients

Notes: Distribution of estimated slope coefficients before (orange) and after (blue) applying the correction factor.

D Additional Tables: Temperature and Crime

Table A.3: Estimated Relationship Between Temperature and Crime: Daily Maximum Temperature

Panel A: Property Crime

	City–Day			County–Day			County–Month		
	IDW (truth)	NN	PRISM	IDW	NN	PRISM	IDW	NN	PRISM
Below 10°F	−1.621***	−1.608***	−1.402***	−1.287***	−1.286***	−1.118***	−2.606**	−2.376**	−2.440**
	(0.351)	(0.352)	(0.312)	(0.289)	(0.286)	(0.265)	(1.091)	(1.022)	(0.998)
10–20°F	−1.310***	−1.319***	−1.172***	−1.097***	−1.100***	−0.977***	−1.356**	−1.540**	−1.592**
	(0.172)	(0.171)	(0.162)	(0.152)	(0.150)	(0.146)	(0.613)	(0.624)	(0.629)
20–30°F	−1.070***	−1.092***	−0.907***	−0.917***	−0.929***	−0.791***	−1.761***	−1.883***	−1.756***
	(0.107)	(0.106)	(0.098)	(0.091)	(0.090)	(0.084)	(0.302)	(0.314)	(0.303)
30–40°F	−0.704***	−0.699***	−0.586***	−0.612***	−0.605***	−0.514***	−1.152***	−1.183***	−1.227***
	(0.069)	(0.070)	(0.063)	(0.059)	(0.059)	(0.053)	(0.202)	(0.210)	(0.216)
40–50°F	−0.339***	−0.343***	−0.264***	−0.296***	−0.296***	−0.234***	−0.747***	−0.698***	−0.692***
	(0.034)	(0.034)	(0.030)	(0.030)	(0.029)	(0.028)	(0.217)	(0.202)	(0.204)
50–60°F	−0.131***	−0.153***	−0.108***	−0.118***	−0.136***	−0.097***	−0.435***	−0.603***	−0.511***
	(0.016)	(0.017)	(0.016)	(0.015)	(0.016)	(0.014)	(0.150)	(0.157)	(0.152)
70–80°F	0.121***	0.113***	0.098***	0.102***	0.097***	0.087***	0.238*	0.180	0.275**
	(0.022)	(0.021)	(0.019)	(0.020)	(0.019)	(0.019)	(0.136)	(0.123)	(0.139)
80–90°F	0.226***	0.217***	0.154***	0.190***	0.185***	0.141***	0.536***	0.481**	0.527***
	(0.050)	(0.047)	(0.045)	(0.045)	(0.045)	(0.043)	(0.201)	(0.192)	(0.202)
Above 90°F	0.277***	0.271***	0.202***	0.214***	0.212***	0.159**	0.857***	0.813***	0.930***
	(0.073)	(0.070)	(0.068)	(0.067)	(0.066)	(0.066)	(0.315)	(0.307)	(0.319)
Obs.	20,155,388	20,155,388	20,155,388	7,636,332	7,636,332	7,636,332	250,920	250,920	250,920

Panel B: Violent Crime

	City–Day			County–Day			County–Month		
	IDW (truth)	NN	PRISM	IDW	NN	PRISM	IDW	NN	PRISM
Below 10°F	−0.275***	−0.279***	−0.265***	−0.225***	−0.227***	−0.213***	−0.128	−0.047	−0.186
	(0.059)	(0.060)	(0.062)	(0.050)	(0.049)	(0.052)	(0.170)	(0.140)	(0.151)
10–20°F	−0.317***	−0.312***	−0.259***	−0.267***	−0.261***	−0.216***	−0.208	−0.243	−0.192
	(0.042)	(0.042)	(0.038)	(0.040)	(0.038)	(0.036)	(0.175)	(0.158)	(0.180)
20–30°F	−0.277***	−0.280***	−0.236***	−0.240***	−0.238***	−0.205***	−0.433***	−0.442***	−0.459***
	(0.026)	(0.026)	(0.025)	(0.025)	(0.025)	(0.023)	(0.073)	(0.086)	(0.084)
30–40°F	−0.210***	−0.208***	−0.173***	−0.181***	−0.179***	−0.150***	−0.276***	−0.285***	−0.292***
	(0.019)	(0.019)	(0.017)	(0.017)	(0.017)	(0.016)	(0.065)	(0.058)	(0.072)
40–50°F	−0.131***	−0.131***	−0.098***	−0.114***	−0.112***	−0.086***	−0.160***	−0.140***	−0.170***
	(0.012)	(0.012)	(0.010)	(0.011)	(0.011)	(0.009)	(0.047)	(0.049)	(0.046)
50–60°F	−0.072***	−0.077***	−0.058***	−0.062***	−0.066***	−0.050***	−0.147**	−0.175***	−0.159***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.058)	(0.052)	(0.055)
70–80°F	0.082***	0.081***	0.062***	0.066***	0.066***	0.051***	0.126**	0.130***	0.137***
	(0.011)	(0.011)	(0.009)	(0.009)	(0.009)	(0.007)	(0.052)	(0.049)	(0.046)
80–90°F	0.143***	0.138***	0.104***	0.114***	0.111***	0.085***	0.129**	0.116**	0.135***
	(0.022)	(0.020)	(0.018)	(0.017)	(0.017)	(0.015)	(0.056)	(0.053)	(0.052)
Above 90°F	0.180***	0.179***	0.137***	0.139***	0.140***	0.104***	0.237***	0.236***	0.271***
	(0.025)	(0.024)	(0.021)	(0.021)	(0.020)	(0.018)	(0.072)	(0.071)	(0.069)
Obs.	20,155,388	20,155,388	20,155,388	7,636,332	7,636,332	7,636,332	250,920	250,920	250,920

Notes: Each panel reports coefficients from variants of equation (5) using crime counts per 100,000 people as the dependent variable. Columns correspond to three levels of aggregation (city–day, county–day, county–month) and three temperature measures (IDW - inverse distance weighting, NN - nearest-neighbor, PRISM). City–day IDW is assumed to be the "truth" benchmark. All specifications include precipitation controls, city/county fixed effects, state-by-month fixed effects, and day-of-week indicators. Standard errors are clustered at the corresponding spatial level and reported in parentheses.

Table A.4: Estimated Relationship Between Temperature and Crime: Daily Mean Temperature

Panel A: Property Crime

	City–Day			County–Day		County–Month	
	IDW	Traditional	PRISM	Traditional	PRISM	Traditional	PRISM
Below 10°F	−1.633***	−1.658***	−1.452***	−1.355***	−1.188***	−2.393***	−2.462***
	(0.243)	(0.251)	(0.225)	(0.215)	(0.198)	(0.773)	(0.782)
10–20°F	−1.293***	−1.312***	−1.110***	−1.104***	−0.949***	−1.541***	−1.553***
	(0.128)	(0.128)	(0.117)	(0.113)	(0.103)	(0.387)	(0.366)
20–30°F	−0.991***	−1.004***	−0.853***	−0.856***	−0.744***	−1.821***	−1.974***
	(0.090)	(0.093)	(0.085)	(0.078)	(0.072)	(0.303)	(0.332)
30–40°F	−0.576***	−0.602***	−0.465***	−0.511***	−0.406***	−1.045***	−0.959***
	(0.057)	(0.059)	(0.051)	(0.050)	(0.044)	(0.209)	(0.201)
40–50°F	−0.273***	−0.287***	−0.228***	−0.243***	−0.203***	−0.698***	−0.805***
	(0.033)	(0.035)	(0.029)	(0.033)	(0.031)	(0.215)	(0.239)
50–60°F	−0.115***	−0.113***	−0.091***	−0.094***	−0.082***	−0.271**	−0.271**
	(0.024)	(0.023)	(0.022)	(0.021)	(0.017)	(0.127)	(0.119)
70–80°F	0.085**	0.083**	0.061	0.069**	0.049	0.114	0.130
	(0.038)	(0.038)	(0.038)	(0.033)	(0.034)	(0.172)	(0.180)
80–90°F	0.141**	0.122*	0.070	0.109*	0.060	0.529**	0.566**
	(0.067)	(0.065)	(0.064)	(0.058)	(0.058)	(0.261)	(0.261)
Above 90°F	0.150	0.074	0.022	−0.081	−0.146	0.156	0.188
	(0.155)	(0.120)	(0.148)	(0.095)	(0.101)	(0.331)	(0.332)
Obs.	20,155,388	20,155,388	20,155,388	7,636,332	7,636,332	250,920	250,920

Panel B: Violent Crime

	City–Day			County–Day		County–Month	
	IDW	Traditional	PRISM	Traditional	PRISM	Traditional	PRISM
Below 10°F	−0.406***	−0.405***	−0.356***	−0.333***	−0.293***	−0.387***	−0.457***
	(0.052)	(0.053)	(0.049)	(0.046)	(0.043)	(0.136)	(0.150)
10–20°F	−0.383***	−0.385***	−0.326***	−0.322***	−0.273***	−0.429***	−0.406***
	(0.035)	(0.035)	(0.032)	(0.033)	(0.030)	(0.115)	(0.100)
20–30°F	−0.333***	−0.331***	−0.289***	−0.280***	−0.247***	−0.497***	−0.560***
	(0.027)	(0.027)	(0.026)	(0.024)	(0.022)	(0.073)	(0.077)
30–40°F	−0.244***	−0.249***	−0.199***	−0.208***	−0.168***	−0.360***	−0.342***
	(0.023)	(0.023)	(0.020)	(0.019)	(0.016)	(0.057)	(0.061)
40–50°F	−0.161***	−0.165***	−0.135***	−0.136***	−0.112***	−0.234***	−0.263***
	(0.017)	(0.017)	(0.015)	(0.014)	(0.012)	(0.047)	(0.049)
50–60°F	−0.089***	−0.090***	−0.070***	−0.072***	−0.056***	−0.162***	−0.144***
	(0.013)	(0.012)	(0.011)	(0.010)	(0.008)	(0.040)	(0.034)
70–80°F	0.057***	0.054***	0.046***	0.043***	0.035***	−0.028	−0.006
	(0.014)	(0.014)	(0.014)	(0.011)	(0.011)	(0.045)	(0.050)
80–90°F	0.097***	0.088***	0.062***	0.069***	0.045***	0.094*	0.126**
	(0.021)	(0.019)	(0.018)	(0.015)	(0.015)	(0.055)	(0.057)
Above 90°F	0.098***	0.090***	0.060*	0.040*	0.010	0.019	0.048
	(0.034)	(0.030)	(0.032)	(0.021)	(0.020)	(0.069)	(0.070)
Obs.	20,155,388	20,155,388	20,155,388	7,636,332	7,636,332	250,920	250,920

Notes: Each panel reports coefficients from variants of equation (5) using crime counts per 100,000 people as the dependent variable. Columns correspond to three levels of aggregation (city–day, county–day, county–month). City–day columns report results using three temperature measures (IDW - inverse distance weighting, Traditional, PRISM), while county–day and county–month columns report results using Traditional and PRISM measures. City–day IDW is assumed to be the "truth" benchmark. All specifications include precipitation controls, city/county fixed effects, state-by-month fixed effects, and day-of-week indicators. Standard errors are clustered at city or county level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Standard errors are clustered at the corresponding spatial level and reported in parentheses.