Appendix for
Fatal Errors: The Mortality Value of Accurate Weather Forecasts

A Data Processing Details

A.1 Primary Estimation Data: Mortality, Weather, and Forecasts

The raw mortality data from the CDC NCHS MCOD files report the day and county of each vital event. All events in a county on a day are added together to generate the county-level number of daily deaths. The deaths are translated into a death rate by dividing by annual county population, as described in Section 3.2.

The day-county structure of the mortality data motivates our processing choices for the PRISM weather data and NDFD forecast data. Both datasets originally provide daily observations on a consistent, high-resolution spatial grid across the U.S. For the forecast data, there are multiple potential observations per day. Forecast models are run and results are reported multiple times per day. For the minimum and maximum temperature forecasts we focus on, the major model runs occur at 12UTC and 00UTC. Based on feedback from National Weather Service meteorologists, we examine the 12UTC forecast, which is available in the early morning for locations in the U.S. and typically forms the basis of the morning forecast on local news. We aggregate the spatial grid to the county level using the following procedure:

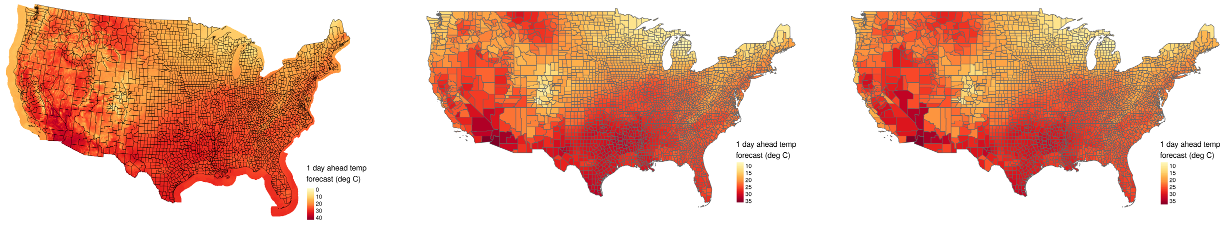
First, for each county, we find the weather and forecast grid points that fall inside the geographic boundary of the county, using 2010 county TIGER/Line shapefile from the Census. Given the high resolution of the underlying datasets (4×4 km for PRISM and either 5×5 or 2.5×2.5 km for NDFD), all counties in our sample contain multiple grid points.

Second, we assign a weight to each grid point based on 2010 population grids from CIESIN (2017). The CIESIN grids are at a roughly 1km resolution, which is higher than either the weather or forecast grid resolutions. Therefore, we use bilinear resampling to reproject the the population grid to match that of the weather and forecast grids.

Third, we calculate population-weighted average values for each weather and forecast observation within the county. The end result is a daily, population-weighted spatial average of the maximum temperature, minimum temperature, total precipitation, dewpoint temperature, maximum temperature forecast for 1 to 6 days ahead, and minimum temperature forecast for 1 to 6 days ahead (the NWS issues forecasts out to 7 days, but given our choice of the 12UTC forecast, the 7-day-ahead minimum temperature forecast is not available). Comparison of an example gridded forecast data and the corresponding county-level data is shown in Figure A1.

Fourth, we correct errors in the forecast data. The NDFD data undergo error checking (such checks are the responsibility of local Weather Forecast Offices), but there are still some identifiable errors in the published data. In particular, from one forecast horizon to the next, there are a small number of observations that have a change in forecast value of

Figure A1: Comparison of Example Raw Gridded Forecast Data and County-level Data


1 day ahead temp forecast (deg C)
0
10
20
30
40
1 day ahead temp forecast (deg C)
10
15
20
25
30
35
1 day ahead temp forecast (deg C)
10
15
20
25
30
35

(a) Gridded Raw Data (b) County Area Average (c) Population-weighted

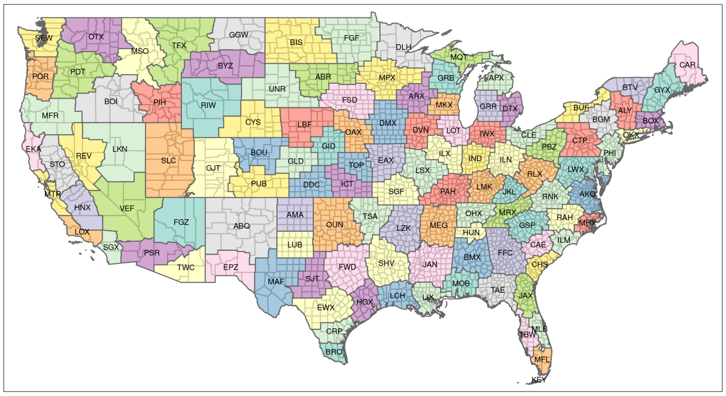
Notes: The maps show the raw, gridded forecast data in panel (a) and the corresponding county-level area and population-weighted average forecasts in panels (b) and (c) respectively. The maps are an example from one day and forecast horizon: the 1-day-ahead forecast for September 9, 2006.

exactly −17.4999°C. When these errors occur, they only appear at one forecast horizon, so we use adjacent forecast horizons to interpolate the erroneous value. In the primary results, we Winsorize the forecast errors, so this data cleaning step has minimal effect on the estimates.

Fifth, we match the timing conventions in the forecast and weather data. The NWS typically uses a noon to noon UTC convention for daily temperature forecasts. Minimum temperatures are forecasted for the nighttime (midnight UTC day to noon UTC day t or 7 p.m day $t-1$ to 7 a.m. day t EST). Maximum temperatures are forecasted for the daytime (noon UTC to midnight UTC). PRISM also typically follows this timing convention, but not as strictly. To match the timing conventions between the two datasets, for maximum and minimum temperatures separately we regress realized temperature on the day t 1-day-ahead forecast and the day $t-1$ 1-day-ahead forecast. For maximum temperature, we find that the day t forecast is sufficient (the day $t-1$ forecast does not predict the realization conditional on the day t forecast). For minimum temperature, we find that both days' forecasts are predictive, with the day t forecast being about twice as predictive as the day $t-1$ forecast. We therefore construct a time-corrected day t minimum temperature forecast that is the weighted average of the original day t and day $t-1$ forecasts with weight 2/3 on the day t forecast and 1/3 on the day $t-1$ forecast. The time-corrected forecast does exhibit forecast sufficiency.

After creating the daily, county-level dataset, we merge counties with identifiers for their NOAA County Warning Area (CWA). CWAs are collections of counties, and the local NWS Weather Forecasting Office (WFO) is responsible for generating forecasts for the CWA. The map of counties and CWAs is shown in Figure A2.

Figure A2: Map of NOAA County Warning Areas (CWAs)



Notes: The map shows (in colored areas with black outlines) the geographic boundaries of County Warning Areas (CWAs), the collection of counties for which a given NWS Weather Forecasting Office is responsible for creating forecasts. State borders are shown in gray, thinner lines. CWAs are typically composed of one or more counties and can cross state borders. There are 116 CWAs in the continental U.S. Some counties are part of multiple CWAs, and in those cases, we assign the county a CWA ID composed of each CWA that it is in. The end result is a many to 1 mapping of all continental U.S. counties to 130 CWAs or CWA groups.

A.2 Air Conditioning

We generate new predictions of air conditioning take-up at the county-by-year level. We begin with individual-level restricted access, biennial American Housing Survey (AHS) data from a Census Research Data Center that contains household-level information on air conditioning (AC) availability and demographic and household information. We link these data with county-by-year climatic characteristics from 1999 to 2020, from Schlenker and Roberts (2009). The AHS sample provides details at the household level that our model uses to predict AC penetration, while partially pooling using other households in the state to improve predictive fit.

We use a multi-step process to select the best model to predict AC availability at the household level. First, we consider all variables shared by the AHS and the public version of the ACS plus weather variables (annual rainfall; annual average temperature; annual maximum temperature; annual minimum temperature; average, max, and min summer temperature; and the annual standard deviation of daily temperature) as possible predictors. The full sample of data is then split, with 1/3 acting as the model testing sample and 2/3 as the hold-out, final prediction sample which helps avoid overfitting or the need for strong sparsity assumptions (Chernozhukov et al., 2018). The samples are blocked to ensure that all states and years are represented in both samples.

Candidate models are evaluated through 5-fold cross validation on the model testing sample, again blocked at the state-year level. Within each cross-validation step, the set of chosen predictors is constructed by taking combinations of the possible predictors and fitting linear multilevel models in R with the `lme4` package (Bates et al., 2015). We model individuals as members of states, so that observations in relatively less sampled states are more strongly pooled toward their group mean to avoid fragile or high-variance out-of-sample predictions.

The first set of candidate models starts by fitting univariate models. We sequentially add each additional potential predictor, keeping the resulting model that achieves the best performance in terms of the Akaike information criterion (AIC). The addition of variables stops if the AIC increases or if a model fails to converge, in which case the last converging model is selected as a candidate. In a second set of models, the set of predictors is chosen through lasso (Tibshirani, 1996) and then a model is fit using `lme4`. These candidate models are then fit with time- and location-varying intercepts and slopes for each combination of up to three of the predictors. The final model is selected from the pool of best-performing models based on out-of-sample predictive fit across the cross-validation folds. The hold-out-sample is then used to estimate the final coefficients for disclosure. We fit the final model on 523,000 observations. We obtain an in-sample root mean squared error (RMSE) of 0.07696 with an average AC penetration rate at the county level of 0.8719.

We then bring these estimated coefficients to individual-level Integrated Public Use Microdata Series (IPUMS USA) data 1% sample household data from the American Community Survey (ACS) from 2005 to 2017. These ACS public use microdata files are geographically identified at the Census Public Use Microdata Area (PUMA). We standardize them by converting them to 2010 PUMA definitions using a 2000 to 2010 PUMA crosswalk by IPUMS. We similarly convert the NOAA climatological data at the county by year level to the PUMA by year level using the Census county to 2010 PUMA crosswalk. We predict household-level AC take up using estimated coefficients from the AHS data applied to our ACS data sample. Finally, we predict AC take-up at the county-by-year level by converting PUMA estimates back to counties: for counties that have a direct match with a PUMA, these estimates are directly applied, and for counties that match to multiple PUMAs, the weighted average of AC take up is estimated for the county-by-year level.

One version of our final data are these direct county-by-year AC take-up estimates, however we note that these unadjusted estimates do not monotonically increase over time as one may theoretically expect. Thus, in a second final estimate, we smooth the county-by-year estimates for each county based on a linear regression of the predicted county-by-year estimates for a given county. We force the overall trend to be (weakly) monotonically increasing. We estimate an average take-up of about 89.2%, broadly consistent with the average take-up of 87.2% in the original AHS data.

A.3 American Time Use Survey

American Time Use Survey data come from the Bureau of Labor Statistics (BLS) and are available at https://www.bls.gov/tus/. The data structure is a repeated cross section. Individuals who have taken part in the CPS are invited to complete a time use diary for a single day's time use. The sample is gathered uniformly throughout the year and across the country. About 10,000 to 12,000 individuals participate each year.

Data are geocoded at a variety of different levels depending on the population density in the location. All observations contain state-level geocoding. For high-density locations, geocoding is at the county level. For intermediate densities, one can use CPS records to geocode the observations at the CBSA level. Details on the geocoding process can be found in Gibson and Shrader (2018). We match individuals to weather and forecast records aggregated to their finest level of geocoding. For clustering, we assign each individual to a WFO either using their county or the WFO that covers the most area in their state.

A.4 Electricity Demand

Electricity demand data come from the US Energy Information Agency (EIA) form EIA-861M and are available at https://www.eia.gov/electricity/data.php. The dataset "Retail sales of electricity to ultimate customers - Monthly" contains monthly, state-level electricity consumption (MWh) and prices (cents/kWh) for residential, commercial, industrial, and other users. We combine the electricity data with weather and forecast data by aggregating the latter to the state-month level. Starting with the county-level data used in the main analysis, we calculate the number of days per month that a state experiences and is forecasted to experience weather in temperature bins (5°C for realized temperature and cold, cool, warm, and hot bins for forecasted temperature, to match Table 1). Rainfall is summed over days in the month and units are converted to meters per month for legibility. Forecast errors are averages of county-level errors. All of these calculations are weighted by county-level population. Finally, we merge the dataset with state-level population from the NIH Surveillance, Epidemiology, and End Results program.

B Additional Theoretical Analysis

B.1 Second-order condition

The following lemma gives sufficient conditions for the second-order condition to hold:

Lemma 2 (Second-Order Condition) *If Assumption 1 holds, $E_{T|f}[h(T, A^*(f))] \leq \epsilon$, and $h_{AA}(T, A^*(T)) \geq 0$, then the second-order condition holds around $A^*(f)$ as ϵ goes to 0.*

Proof. Differentiating the right-hand side of (2) with respect to A, applying Assumption 1 and Lemma 1, and letting $E_{T|f}[h(T, A)]$ be small, the second-order condition holds around $A^*(f)$ if $-C''u' + [C']^2u'' - E_{T|f}[h_{AA}](u - v) < 0$, which holds if $E_{T|f}[h_{AA}] \geq 0$. ■

Given that an individual's daily mortality risk is not generally large, the second-order condition should hold in our empirical application as long as the hazard function is not too concave in actions.

B.2 Identifying properties of the hazard function

Here we show how estimating which type of adaptation environment holds is informative about properties of the hazard function.

Begin by establishing that actions are sensitive to forecasts.

Assumption 3 (Constant VSL) *Around forecast f, the VSL is constant.*

Lemma 3 (Actions Respond to Forecasts) *If Assumptions 1 through 3 hold and the second-order condition for optimality of actions holds when the forecast is f, then* $\lim_{\epsilon \to 0} A^{*\prime}(f) \propto -h_{AT}(f, A^*(f))\, VSL$.

Proof. Applying the implicit function theorem to (4) and using the second-order condition, $A^{*\prime}(f) \propto -\frac{\partial E_{T|f}[h_A((T,A^*(f))]}{\partial f} VSL(f) - E_{T|f}[h_A((T, A^*(f))]\, VSL'(f)$. Using (5) and $VSL'(f) = 0$ (and thus dropping the argument of VSL), this becomes

$$A^{*\prime}(f) \propto -\left\{ h_{AT}(f, A^*(f)) + \frac{1}{2} h_{ATTT}(f, A^*(f)) Var[T|f] + \frac{1}{2} h_{ATT}(f, A^*(f)) \frac{\mathrm{d}Var[T|f]}{\mathrm{d}f} \right\} VSL.$$

The lemma follows from applying Assumptions 1 and 2. ■

Actions are sensitive to forecasts when the marginal effect of temperature on mortality risk depends on the actions chosen (i.e., when $h_{AT} \neq 0$).

Now consider what we learn about the hazard function:

Lemma 4 *If Assumptions 1 through 3 hold at $f = T$ with ϵ arbitrarily small and the second-order condition for optimality of actions holds when the forecast is $f = T$, then adaptation is appropriate if and only if* $\lim_{\epsilon \to 0} h_{AA}(T, A^*(T)) > 0$.

Proof. Using Lemmas 1 and 3, equation (6) implies that $\lim_{\epsilon \to 0} \; \mathrm{d}^2 h(T, A^*(T+e))/\,\mathrm{d}e^2|_{e=0} > 0$ if and only if $\lim_{\epsilon \to 0} h_{AA}(T, A^*(T)) > 0$. And from Definition 1, adaptation is appropriate if and only if $\mathrm{d}^2 h(T, A^*(T+e))/\,\mathrm{d}e^2|_{e=0} > 0$. ■

If we estimate that mortality risk is convex in forecast errors, then we can use Lemma 4 to conclude that the hazard is convex in actions around accurate forecasts. But if we instead estimate that mortality risk is not convex in forecast errors, then we can conclude that the hazard is linear or concave in actions around accurate forecasts.

B.3 Proof of Proposition 2

Second-order approximate $V(f)$ around $f = T$ inside $\bar{V}(T)$:

$$\bar{V}(T) \approx V(T) + \frac{1}{2} V''(T) \, Var[e|T]. \tag{A-1}$$

Therefore:

$$\frac{\mathrm{d}\bar{V}(T)}{\mathrm{d}Var[e|T]} \approx \frac{1}{2} V''(T), \tag{A-2}$$

with the approximation becoming exact when Assumption 1 holds with ϵ small. Second-order approximating $h(f-e, A^*(f))$ around $e = 0$ yields:

$$E_{e|f}[h(f-e, A^*(f))] \approx h(f, A^*(f)) + \frac{1}{2} h_{TT}(f, A^*(f)) \, Var[e|f],$$

with the approximation again becoming exact when Assumption 1 holds with ϵ small. Substitute into $V(T)$ and apply Assumption 2 in order to hold $Var[e|T]$ constant in a neighborhood of T:

$$\begin{aligned} V'(T) = & - \left[h_T(T, A^*(T)) + \frac{1}{2} h_{TTT}(T, A^*(T)) Var[e|T] \right] \, [u(w - C(A^*(T))) - v(w - C(A^*(T)))] \\ & + \frac{\mathrm{d}V(T)}{\mathrm{d}A} A^{*\prime}(T). \end{aligned}$$

Differentiating again yields:

$$\begin{aligned}
V''(T) = &- \left[h_{TT}(T, A^*(T)) + \frac{1}{2} h_{TTTT}(T, A^*(T)) Var[e|T]\right] [u(w - C(A^*(T))) - v(w - C(A^*(T)))] \\
&- A^{*\prime}(T) \left[h_{AT}(T, A^*(T)) + \frac{1}{2} h_{ATTT}(T, A^*(T)) Var[e|f]\right] \\
&\quad [u(w - C(A^*(T))) - v(w - C(A^*(T)))] \\
&+ A^{*\prime}(T)\, C'(A^*(T)) \left[h_T(T, A^*(T)) + \frac{1}{2} h_{TTT}(T, A^*(T)) Var[e|T]\right] \\
&\quad [u'(w - C(A^*(T))) - v'(w - C(A^*(T)))] \\
&+ \left.\frac{\mathrm{d}\frac{\mathrm{d}V(f)}{\mathrm{d}A} A^{*\prime}(f)}{\mathrm{d}f}\right|_{f=T}.
\end{aligned}$$

The final line vanishes because the first-order condition must hold at all f. Substitute from (4) and then from (5),

$$\begin{aligned}
V''(T) \approx &- \left[h_{TT}(T, A^*(T)) + \frac{1}{2} h_{TTTT}(T, A^*(T)) Var[e|T]\right] [u(w - C(A^*(T))) - v(w - C(A^*(T)))] \\
&- A^{*\prime}(T) \left[h_{AT}(T, A^*(T)) + \frac{1}{2} h_{ATTT}(T, A^*(T)) Var[e|f]\right] \\
&\quad [u(w - C(A^*(T))) - v(w - C(A^*(T)))] \\
&- A^{*\prime}(T) \left[h_A(f, A^*(f)) + \frac{1}{2} h_{ATT}(f, A^*(f)) Var[T|f]\right] VSL(T) \\
&\quad \left[h_T(T, A^*(T)) + \frac{1}{2} h_{TTT}(T, A^*(T)) Var[e|T]\right] [u'(w - C(A^*(T))) - v'(w - C(A^*(T)))].
\end{aligned}$$

Using Assumption 1 and Lemma 1,

$$\lim_{\epsilon \to 0} V''(T) = - \lim_{\epsilon \to 0} \left[h_{TT}(T, A^*(T)) + A^{*\prime}(T)\, h_{AT}(T, A^*(T))\right] [u(w - C(A^*(T))) - v(w - C(A^*(T)))].$$

Observe that

$$\frac{\mathrm{d}}{\mathrm{d}e} \frac{\partial h(T, A^*(T + e))}{\partial T} = h_{AT}(T, A^*(T + e))\, A^{*\prime}(T + e),$$

which, from Lemma 3, is strictly negative if Assumption 3 and the second-order condition for optimality of actions both hold when the forecast is f. We then have:

$$\lim_{\epsilon\to 0} V''(T) = -\lim_{\epsilon\to 0}\left[\left.\frac{\partial^2 h(T, A^*(T+e))}{\partial T^2}\right|_{e=0} + \left.\frac{\mathrm{d}}{\mathrm{d}e}\frac{\partial h(T, A^*(T+e))}{\partial T}\right|_{e=0}\right] \left[u(w - C(A^*(T))) - v(w - C(A^*(T)))\right].$$

Substituting into (A-2), we find:

$$\lim_{\epsilon\to 0}\frac{\mathrm{d}\bar{V}(T)}{\mathrm{d}Var[e|T]} \approx -\lim_{\epsilon\to 0}\frac{1}{2}\left[\left.\frac{\partial^2 h(T, A^*(T+e))}{\partial T^2}\right|_{e=0} + \left.\frac{\mathrm{d}}{\mathrm{d}e}\frac{\partial h(T, A^*(T+e))}{\partial T}\right|_{e=0}\right] \left[u(w - C(A^*(T))) - v(w - C(A^*(T)))\right]. \tag{A-3}$$

Using (A-1), observe that, under Assumption 1,

$$\lim_{\epsilon\to 0}\frac{\mathrm{d}\bar{V}}{\mathrm{d}w} = E_{T|f}[1 - h(T, A)]\, u'(w - C(A^*(T))) + E_{T|f}[h(T, A)]\, v'(w - C(A^*(T))). \tag{A-4}$$

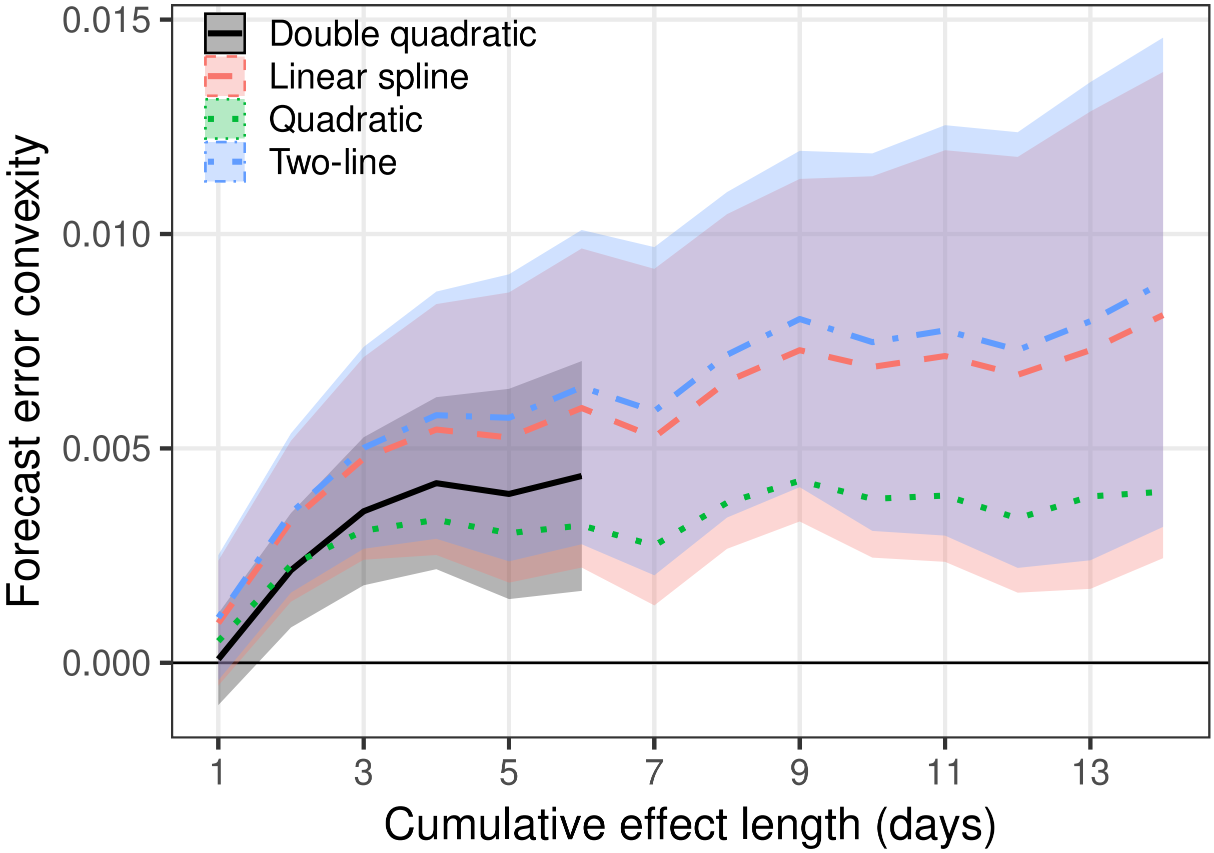
Plugging (A-3) and (A-4) into the definition of $WTP(T)$ in (8) and substituting from (3) yields the expression in the proposition.

C Testing Assumptions Underlying Forecast Benefit Estimates and Robustness to Functional Form

Section 6 shows the benefit of forecast improvements based on the 6-day cumulative effect of forecasts on mortality. One assumption underlying those results is that the effect of forecasts on mortality is persistent. Previous research on the effect of realized temperature on mortality has shown that there can be dynamic effects over the days following a temperature realization (Deschênes and Moretti, 2009, Heutel et al., 2021). Figure A3 shows the average marginal effect across the full temperature distribution of a more erroneous forecast (the same value reported in Column 5, row 3 of Table 3) using a two-week distributed lag model.[41] The results show that on the day the forecast arrives, the marginal effect is roughly 0.001 deaths per 100,000 people. This rises to around 0.002 after just one additional day, continues to rise through day 3, then stays roughly stable after that point. The value is significantly positive over all but the first day. This stability of the estimate supports our assumption of a persistent effect on mortality.

[41] Examining longer ranges of cumulative effects is computationally infeasible using a standard distributed lag model given the high dimensionality of the estimating equation.

Figure A3: Testing Counterfactual Persistence Assumption: Cumulative Effects



Notes: Shows estimates of the cumulative average convexity of mortality with respect to forecast error based on regressions estimated using Equation (10) ("Two-line"), a variant of that equation using a linear spline ("Linear spline"), a variant using a quadratic ("Quadratic"), and the baseline value estimating equation (14) ("Double quadratic"). All are fit to the baseline data and use 14 lags except the double quadratic. The shaded area shows the 95% confidence interval based on standard errors clustered at the CWA level.

A second assumption underpinning Table 3 is that the estimated forecast-mortality relationship holds under the counterfactual forecast. We generate descriptive evidence on this assumption by looking at how the effect varies by average forecast quality (measured by RMSE) in the sample. Results are shown in Figure 4. These estimates are non-causal, as locations with more accurate forecasts may differ from locations with less accurate forecasts in unobserved ways.[42] In each forecasted temperature bin, forecasts are more valuable in locations that have more accurate forecasts on average. If anything, this result suggests that our estimates are lower bounds on the mortality benefit of improved forecasts because more accurate forecasts are associated with stronger responses to forecasts.

For robustness, we assess the different estimates of lives saved that we find when estimating with different functional forms. In particular, we use the estimates of equation (10) reported in Table 1, a version that uses a linear spline rather than the two-line formulation, and a quadratic over forecast errors but not over temperature.[43]

The cumulative convexity derived from each of these equations is shown in Figure A3. The figure shows the average convexity across all temperature bins. For each functional form, the convexity rises for about 4 days then stabilizes. For the more parsimonious functional forms that permit longer cumulative effects to be computationally tractable, this stability is exhibited through two weeks with no indication of changing. Overall, the two linear models (two-line and linear spline) exhibit the largest average convexity. The two quadratic models exhibit slightly smaller average convexity, though still within the confidence bands of the linear models.

D Sensitivity to Breakpoint Choice for Two-line Test

The Simonsohn (2018) two-line procedure involves first selecting a breakpoint then fitting an interrupted regression on either side of that breakpoint. For selecting the breakpoint, Simonsohn proposes the "Robin Hood algorithm." The algorithm "donates" points from the more precisely estimated side of the interrupted regression to the weaker side so that overall power of the test is improved. In the original algorithm from Simonsohn (2018), the researcher first fits a flexible function relating the left-hand and right-hand side variables. This fit can be done using a nonparametric or semiparametric procedure. The researcher then

[42] There could be omitted variables that could cause forecasts to be more accurate and also cause errors to matter more (e.g., experiencing hot weather more frequently), and there could be selection in which places get more accurate forecasts (e.g., radars or skilled meteorologists may be directed to places where the National Weather Service believes that forecasts are more valuable). We adjust for observable confounders by including them in the regression, but unobservables might still confound the relationship.

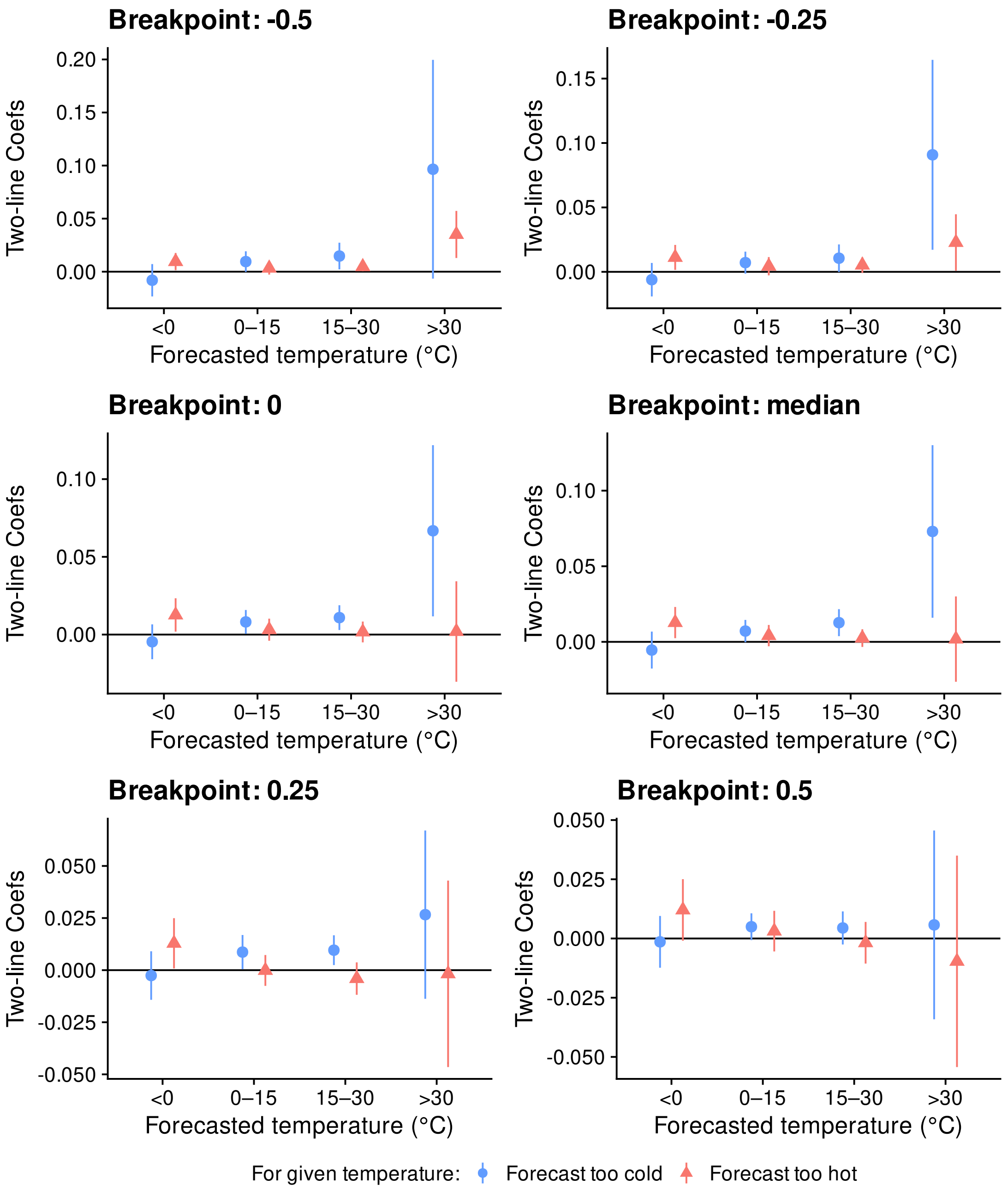
[43] We also estimated net benefits using a flexible polynomial specification rather than bins of forecasted temperature. The overall WTP estimates are highly similar in both cases, and the results are available upon request.

generates fitted values and chooses an initial candidate breakpoint, x_0, based on the most extreme fitted value. A range of additional candidate breakpoints is chosen by considering all x values within 1 standard error of the most extreme fitted value. This set of candidate breakpoints has upper bound x_H and lower bound x_L.

Next, the researcher fits an interrupted regression using the initial, candidate breakpoint x_0. The interrupted regression fit will lead to two initial slope estimates with associated standard errors $\hat{s}_H$ and $\hat{s}_L$ for the estimate above and below the breakpoint respectively. The final breakpoint is chosen by shifting the breakpoint to increase the number of points given to the less precisely estimated slope. Simonsohn proposes that the final breakpoint be $(\hat{s}_{imp}/(\hat{s}_L + \hat{s}_H) * 100$ percent of the way toward the edge of the boundary of candidate breakpoints, where $\hat{s}_{imp}$ is the standard error of the relatively less precise estimate.

In our setting, we have strong *a priori* reasons for preferring a breakpoint around 0 or median error, based on the theoretical analysis in Section 2. We thus use a median breakpoint for our main results. Figure A4 shows the estimated marginal effects using a range of different breakpoints. The breakpoint is indicated in the title of each panel. The results are very similar to the baseline (median breakpoints) when using a breakpoint of 0 because the median forecast error is close to 0 for all forecasted temperature bins. Results are close to the baseline results for the moderately negative and positive breakpoints (-0.25 and 0.25) although the "forecast too cold" coefficient in the hot temperature bin is not significant in the latter case. Results are weakest when the breakpoint is 0.5 (bottom right panel), a point that is half of a standard deviation away from the median error in most bins.

Figure A4: Robustness and Sensitivity Checks



Notes: The figures show sensitivity to the choice of breakpoint in estimating Equation 10. All models use the same functional form, lag length, and controls as the baseline results. The lines are 95% confidence intervals based on standard errors clustered at the CWA level. The panel titles indicate the forecast error breakpoint used in the panel. The baseline results are those that use median breakpoints and correspond to the results in Table 1.

E Additional Robustness Checks

Figure A5 shows robustness and sensitivity checks for the main results shown in Table 1. The red triangles show the marginal effect of a forecast that is too hot and the blue circles show the marginal effects for a forecast that is too cold. The lines are 95% confidence intervals. The figures show results for forecasted cold temperatures (panel a), cool temperatures (panel b), warm temperatures (panel c), and hot temperatures (panel d). Each panel is broken into 3 sections. The first section varies the standard error clustering; the second section varies the non-weather related controls; and the third section varies the realized weather, pollution, and forecast controls.

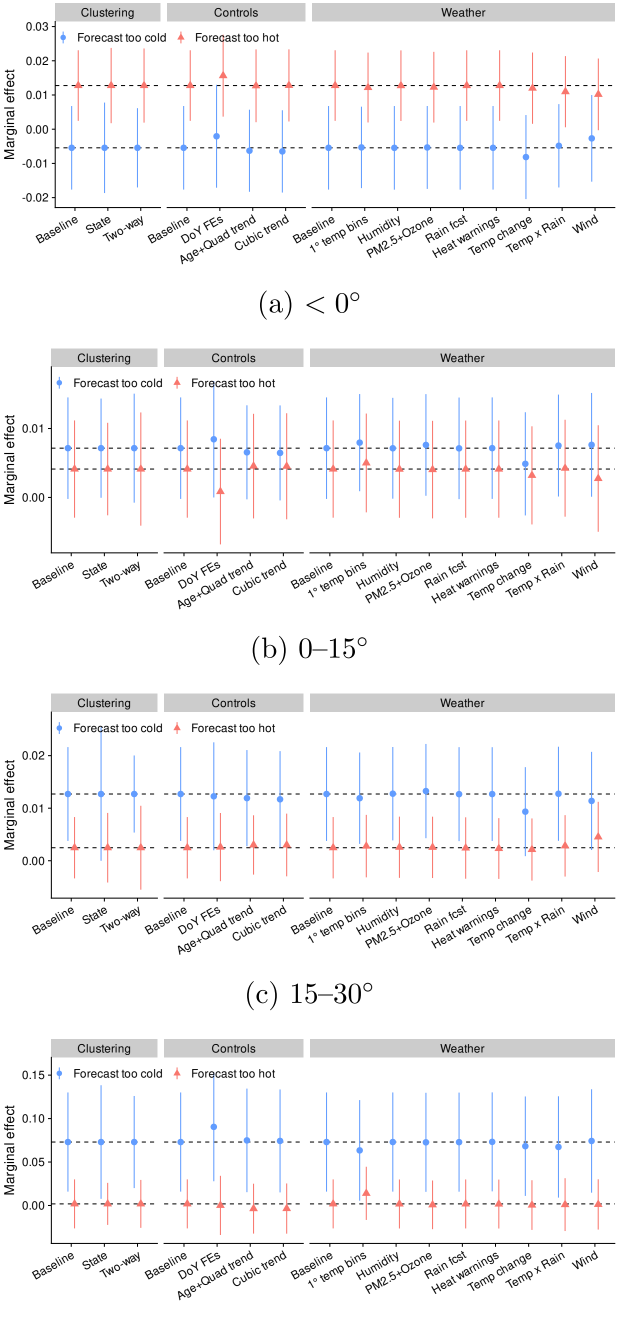
In the clustering section, the baseline estimate is clustered at the CWA level. "State" clusters at the state, and "Two-way" clusters at the county and year level. The results are typically similar across all of these clustering schemes.

In the controls section, baseline estimate includes controls for 5°C bins of realized temperature, lags of indicators for above median precipitation, date fixed effects, and county-by-month fixed effects interacted with linear time trends. "DoY FEs" replaces the month fixed effects with day-of-year fixed effects. "Age+Quad trend" add a quadratic year trend interacted with county-by-month fixed effects and month fixed effects interacted with four population age indicators. This exactly matches the control set used in Barreca et al. (2016) but adapted to our county-level dataset rather than a state-level dataset. "Cubic trend" adds cubic trends interacted with county-by-month fixed effects. The day of year fixed effects have the strongest effect on the estimates among this set of checks.

The third section varies controls for weather, pollution, and other forecasts. Unless otherwise stated, the controls for weather are included using 8 bins equally spaced using percentiles of the data. "1° temp bins" replaces the 5°C bins with finer controls for realized temperature. "Humidity" controls for the relative humidity. "PM2.5+Ozone" controls for county-level ambient ozone and $PM_{2.5}$ concentrations from EPA's RSIG Fused Air Quality Surface Using Downscaling (FAQSD) files available here: https://www.epa.gov/hesc/rsig-related-downloadable-data-files. "Rain fcst" includes 1-day-ahead rainfall forecasts binned to match the realized rainfall controls, "Temp change" includes the change in temperature between day t and $t-1$, "Temp x Rain" interacts the realized rain and temperature controls, and "Wind" controls for both direction and speed of wind as measured by NOAA's North American Regional Reanalysis (NARR) available here: https://psl.noaa.gov/data/gridded/data.narr.html.

Like PRISM, the NARR dataset combines individual weather observations with a model (in this case, the NCEP Eta weather model) to produce weather measures on a consistent grid across the U.S (Mesinger et al., 2006). The grid has a spatial dimension of roughly 32km,

Figure A5: Robustness and Sensitivity Checks



(a) < 0°

(b) 0–15°

(c) 15–30°

(d) > 30°

Notes: The figures show robustness and sensitivity checks on the main results reported in Section 4. All models use the same functional form and lag length as the baseline results. The lines are 95% confidence intervals based on standard errors clustered at the CWA level. The labels indicate the change or addition. For comparison, "Baseline" reproduces the baseline estimate, with controls described in Section 4.1. In all cases, added weather variables are controlled for non-parametrically using quantile bins, and 6 lags of the bins are included to match the lag length of the forecast error. Note that wind and rainfall forecasts are only available for a subset of the observations, so the sample changes.

and we take a spatial average of the values in each county to match our estimation sample.[44] Because the wind data grid is coarser than the PRISM grid, we lose some county observations when we include wind. The sample also changes for the rainfall point forecast because rainfall point forecasts were included in the NDFD at a later date than the temperature point forecasts used in the main analysis.

Among the weather controls, the largest effects come from using finer temperature bins, including temperature changes, and including wind. Finer temperature bins reduce the effect of negative forecast errors during hot periods and increase the effect of forecast errors that are too hot. Temperature changes reduce the effect of negative forecast errors during both cool and warm periods. Including wind moves marginal effects toward each other during both cold and warm periods.

F Counterfactual Approximation Quality

In Section 2, we derive marginal conditions for forecast value. These conditions are second-order approximations to the value of a change in forecast error distribution. For a marginal change in forecast error standard deviation, this follows from Proposition 1,

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\mathrm{d}\bar{h}(T)}{\mathrm{d}\sigma_{e|T}}n(T) &= \lim_{\epsilon\to 0}\frac{\mathrm{d}\bar{h}(T)}{\mathrm{d}Var[e|T]}\frac{\mathrm{d}Var[e|T]}{\mathrm{d}\sigma_{e|T}}n(T)\\ &= \lim_{\epsilon\to 0}\left.\frac{\mathrm{d}^2 h(T, A^*(T+e))}{\mathrm{d}e^2}\right|_{e=0}\sigma_{e|T}n(T) \qquad \text{(A-5)}\end{aligned}$$

where $\sigma_{e|T}$ is the standard deviation of forecasts (and forecast errors) given temperature T. For a $X \times 100$ percent reduction in forecast error standard deviation, the change in mortality is scaled by $(1-(1-X)^2)\sigma_{e|T}^2$ instead of $2\sigma_{e|T}$.

For a discrete change in the forecast error distribution, the value is given by the difference in expected value under the counterfactual and actual distributions. The approximation will be accurate if the distribution of errors is close to normal or if the mortality hazard function is approximately quadratic in forecast errors. The approximation is practically useful because it is faster to compute. Table A1 compares the estimated counterfactual across all realized temperatures using both the approximation and a nonparametric estimate. One can see that the approximation accurately reproduces the results from the nonparametric estimator in this setting, with a difference in estimates of no more than 5%.

[44] Further details on the steps we follow to process the wind data can be found in Missirian (2020).

Table A1: Comparison of Counterfactual Approximation and Nonparametric Calculation for Monetized Lives Saved

	(1)	(2)	(3)	(4)
Forecasted temperature:	$< 0°$	$0 - 15°$	$15 - 30°$	$> 30°$
Approximation method				
Nonparametric	110.923	768.558	1268.245	44.925
Approximation	116.476	789.543	1290.889	46.356
Approx./Nonpar.	1.05	1.027	1.018	1.032

Notes: The table compares estimates of the counterfactual lives saved from a 50% improvement in forecast error calculated using a nonparametric approximation (row 1) and second-order approximation (row 2). The counterfactual values correspond to the "50% improvement" row from Table 3.

G Heterogeneity by Demographics, Cause of Death, and Region

The CDC mortality records provide three dimensions of demographic information about the deceased individuals. They also list the cause of death. Figures A6a, A6b, A6c, and A7 show heterogeneity results along these different dimensions, based on estimates of equation (14) where the left-hand side variable has been replaced with mortality for the demographic or cause of death group listed in the figure. In all panels, the y-axis shows the annual lives saved per 100,000 people. In each of the figures, the top left panel shows the estimates for cold forecasted temperatures, the top right panels show it for cool forecasted temperatures, bottom left for warm, and bottom right for hot.

For sex (Figure A6a), the coefficients are almost always the same for both men and women. The one exception is forecasts that are too cold on days that are forecasted to be cold. This reduces mortality for women, if anything. This results in per capita lives saved that are similar for men and women at all but cold temperatures, where men have more positive benefit while women have a value near zero (although the difference is not statistically significant).

For age (Figure A6b), the strongest effects come from individuals older than 35. Point estimates are small for young people. In general, there is monotonically increasing per capita lives saved from forecast improvements as individuals get older for all temperature bins. The one exception is the cold bin, where individuals between ages 75 and 84 experience the highest per capita mortality reductions and all other groups show mortality reductions near zero. In unreported results, there is substantial heterogeneity in the effect within the 0 to 19 age group, with the largest point estimates for children between 1 and 5 years old, and a slightly negative point estimate for infants less than 1. In all cases, however, the confidence intervals for these groups are extremely wide.

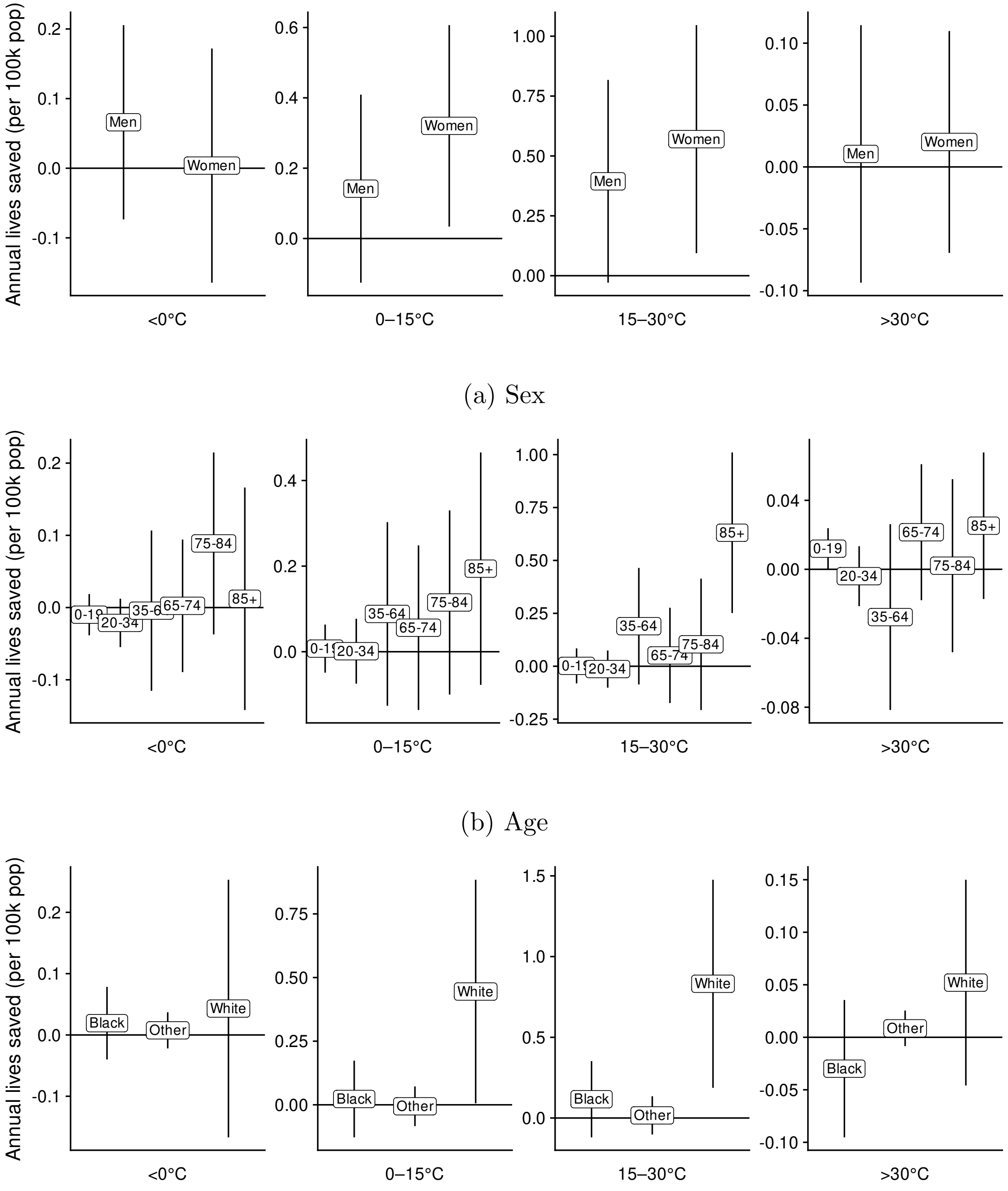
Figure A6c compares forecast effects by the race of the deceased, with separate estimates for individuals who identified as white, Black, or other races (Asian or Pacific Islander, and American Indian or Alaska Native). The effect of forecasts on mortality is substantially greater for white individuals than for all other individuals. Notably, in Figure A6c (the same estimates as reported in the body of the text in Figure 5), the point estimates indicate that forecasts across the temperature distribution have close to zero effect on mortality for all people of color.

Figure A7 shows estimates by cause of death. Causes of death reported on death certificates are subject to discretion by the individual filling out the death certificate, so all results should be taken as noisy and weakly informative. In terms of point estimates, the causes that are most significantly associated with forecasts are acute respiratory failure, accidents, cardiovascular disease, as well as other disease and “all other” which captures any cause of death that is not explicitly categorized.

The strong associations with respiratory, cardiovascular, and cancer deaths is consistent with the findings on leading causes of death from temperature exposure (Deschênes and Moretti, 2009). The higher association with accidents is not found in studies of realized temperature and could be due to avoidance behavior engaged in by individuals to try to reduce their exposure to extreme weather.

The Figures A8 and A9 show the annual, lives saved per 100,000 people from marginal forecast improvements for the 9 NOAA climate regions (indicated on the x-axis of each figure). The points are estimates and the whiskers are 95% confidence intervals.

Figure A6: Heterogeneity by Demographics of the Deceased



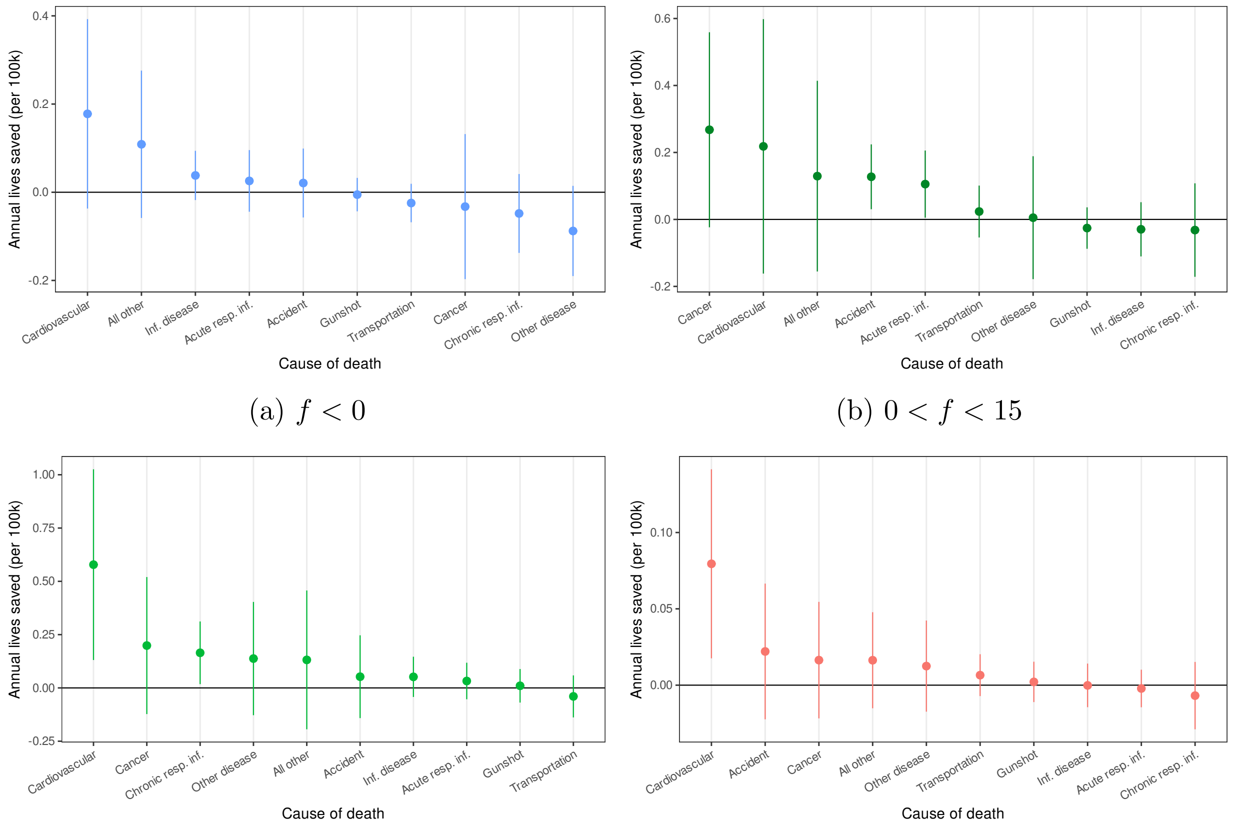
(a) Sex

(b) Age

(c) Race

Notes: The figure shows the 6-day cumulative annual, lives saved per 100,000 people from a marginal decrease in 1-day-ahead forecast error. Estimates for each demographic category come from a separate model fit using Equation (14) on the baseline data where the dependent variable is mortality in the indicated demographic group. The range of forecasted temperature is indicated below each figure. Demographic categories are shown in the labelled points, which are also the point estimates derived from estimation. The lines are 95% confidence intervals based on standard errors clustered at the CWA level.

Figure A7: Heterogeneity by Cause of Death



(a) $f < 0$

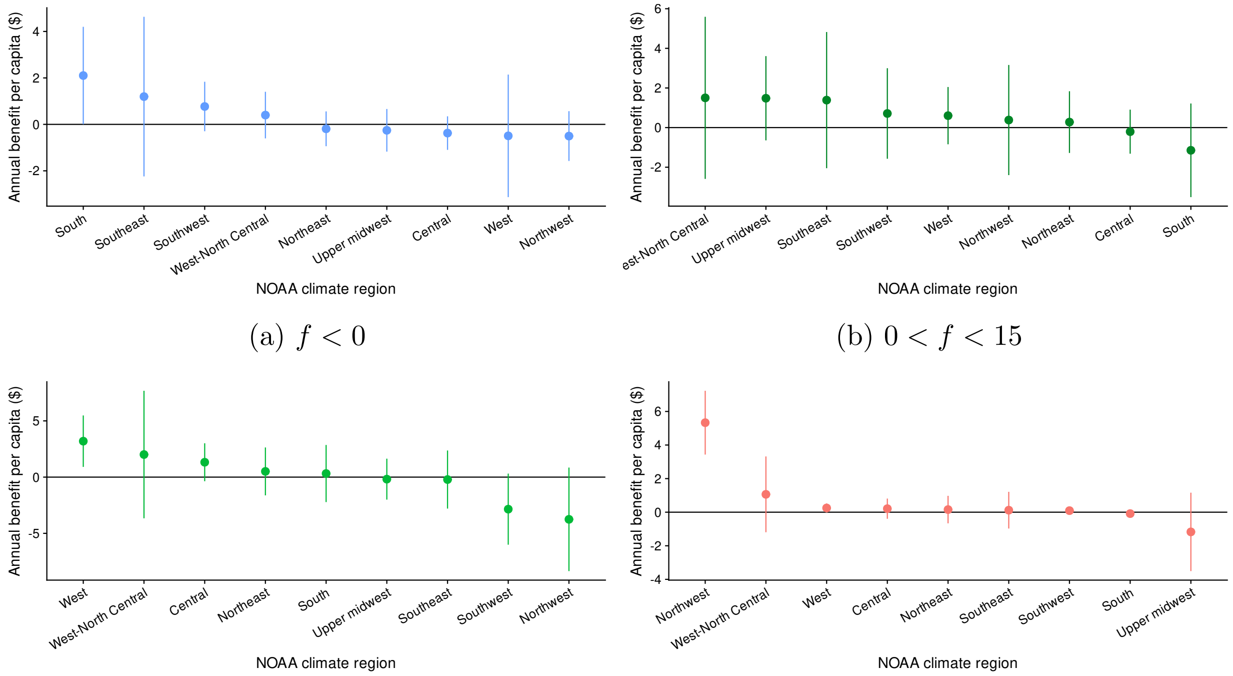
(b) $0 < f < 15$

(c) $15 < f < 30$

(d) $30 > f$

Notes: The figure shows the 6-day cumulative effect on annual lives saved per 100,000 people from a marginal reduction in forecast error. Estimates for each cause of death come from a separate model fit using Equation (10) on the baseline data. The cause of death is shown on the x-axis, and estimates are ordered by the effect size for negative forecast errors for $f < 0°\text{C}$ and for positive forecast errors for all other panels. The range of forecasted temperature is indicated below each figure. Circles are the point estimates and the lines are 95% confidence intervals based on standard errors clustered at the CWA level.

Figure A8: Monetized Mortality Benefits of Marginal Forecast Improvement: Regional Heterogeneity by Forecasted Temperature



(a) $f < 0$

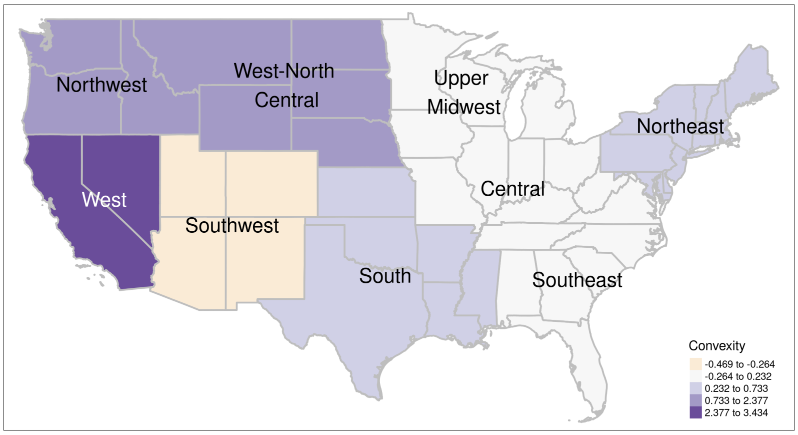
(b) $0 < f < 15$

(c) $15 < f < 30$

(d) $30 > f$

Notes: The figure shows the 6-day cumulative percent increase in the per-capita mortality rate, monetized using the VSL, from a 1°C increase in forecast mean absolute error based on a single model fit using Equation (10) on the baseline data. The forecast error variables are interacted with indicators for each NOAA climate region in the Continental U.S. The circles are the point estimates and the lines are 95% confidence intervals based on standard errors clustered at the CWA level. A map showing these effects spatially can be found in Figure A9.

Figure A9: Spatial Heterogeneity in Forecast Effect



Notes: The figure shows the 6-day cumulative effect on the mortality rate from a 1°C reduction in forecast absolute error based on Equation (10) estimated across NOAA climate regions. Darker purple colors indicate stronger benefits from forecast improvements, while lighter, orange colors indicate lower benefits.

H Longer-horizon Forecasts

The NWS issues point forecasts with horizons of up to 1 week. Table A2 shows 2-day cumulative effects when including both the 1-day-ahead forecast and longer-horizon forecasts in the estimation simultaneously. We focus on 2-day effects to simplify the interpretation when including multiple forecast horizons—looking only over 2 days means that the 1-day-ahead forecast is always the most recent, available information included in the regression. If there are adjustment costs that hamper individuals from acting on shorter-horizon forecasts, then longer-horizon forecasts can provide more adaptation benefits. This will show up as a convex relationship between mortality and the longer-horizon forecast, even conditional on the shorter-horizon forecast.

Table A2: Effects by Forecast Horizon: Convexity of Pooled Estimates

	(1)	(2)	(3)
	Mortality rate	Mortality rate	Mortality rate
1-day ahead convexity	0.0030***	0.0022***	0.0022**
	(0.0006)	(0.0006)	(0.0008)
3-day ahead convexity		0.0008**	
		(0.0004)	
6-day ahead convexity			0.0008***
			(0.0002)
Dependent var. mean	2.25	2.24	2.24
N	13,529,776	13,408,395	11,078,870
N Clusters	130	130	130

Notes The table shows 2-day cumulative effects from estimation of versions of Equation (10) that also include longer-horizon forecasts and use a quadratic specification rather than a two-line specification to capture non-linearity in the effect of forecast errors on mortality. The dependent variable is the daily mortality rate per 100,000 people. The coefficients are the average marginal effect of a more erroneous forecast, pooled across all forecasted temperature bins. All models include the baseline model covariates and weighting. Standard errors, clustered at the CWA-level, are below each estimate. Significance: $p < .10$, ** $p < .05$, *** $p < .01$.

The results in Table A2 are consistent with longer-horizon forecasts providing additional value over-and-above the day-ahead forecasts. For both the 3- and 6-day forecasts, forecast errors have a significant, convex relationship with mortality. Notably, in Columns (2) and (3), the sum of the effects of the 1-day-ahead forecast and the 3- or 6-day ahead forecast are approximately equal to the effects of the 1-day-ahead forecast in Column (1), where the longer-horizon forecasts are not included. Forecasts at all horizons are highly correlated, so the 1-day-ahead forecast effect in Column (1) (and in our other results) captures many of the benefits of all horizons of forecasts issued by the NWS.

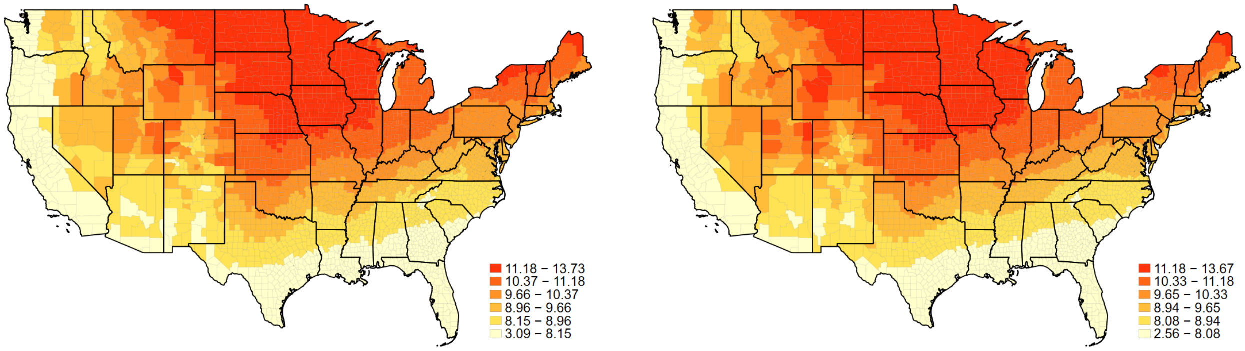
I Additional Figures and Tables

Table A3: Summary Statistics

Variable	Mean	S.D.	Observations
Daily all-cause mortality rate (per 100,000)	2.25	1.73	13,699,990
Average temperature (°C)	14.557	10.052	13,699,990
1-day-ahead avg. temperature forecast (°C)	14.515	9.964	13,699,990
1-day-ahead forecast error (°C)	-0.041	1.146	13,699,990
1-day error if $f < 0$ (°C)	-0.008	1.396	1,775,001
1-day error if $0 \leq f < 15$ (°C)	-0.007	1.229	5,261,602
1-day error if $15 \leq f < 30$ (°C)	-0.075	1.031	6,516,717
1-day error if $f \geq 30$ (°C)	0.060	0.905	146,670

Notes: The table shows summary statistics for the primary variables in the estimation sample, weighted by county population. The difference between average realized temperature and average forecasted temperature does not necessarily equal the average forecast error due to rounding.

Figure A10: Unconditional Variation in Temperature and Day-ahead Forecast

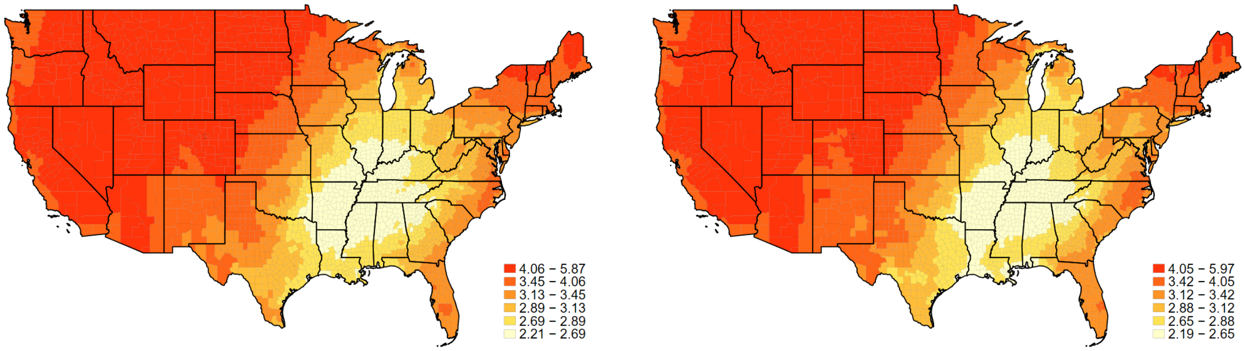


(a) St. dev. of average temperature

(b) St. dev. of 1-day-ahead forecast

Notes: The maps show the standard deviation of the unconditional average temperature (left panel) or the 1-day-ahead forecast of average temperature (right panel). For an indication of the identifying variation conditional on controls, compare these maps to the maps in Figure A11.

Figure A11: Residual Variation in Temperature and Day-ahead Forecast

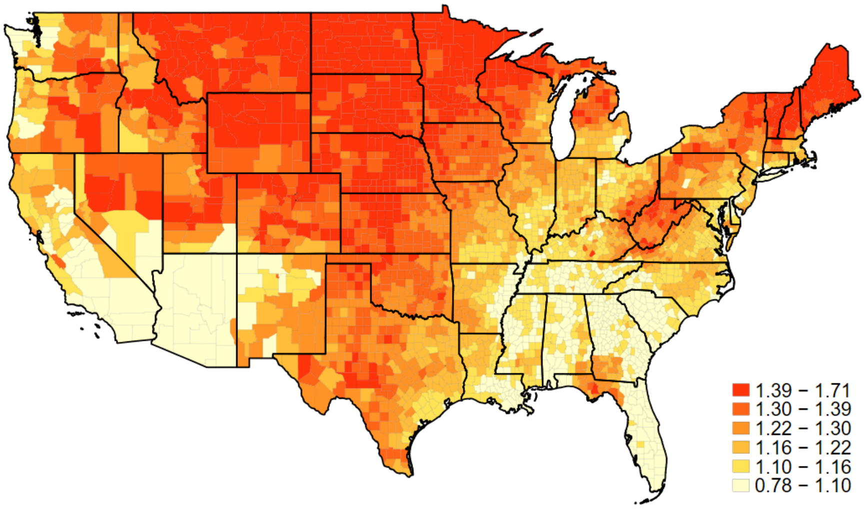


(a) St. dev. of residual average temperature

(b) St. dev. of residual 1-day-ahead forecast

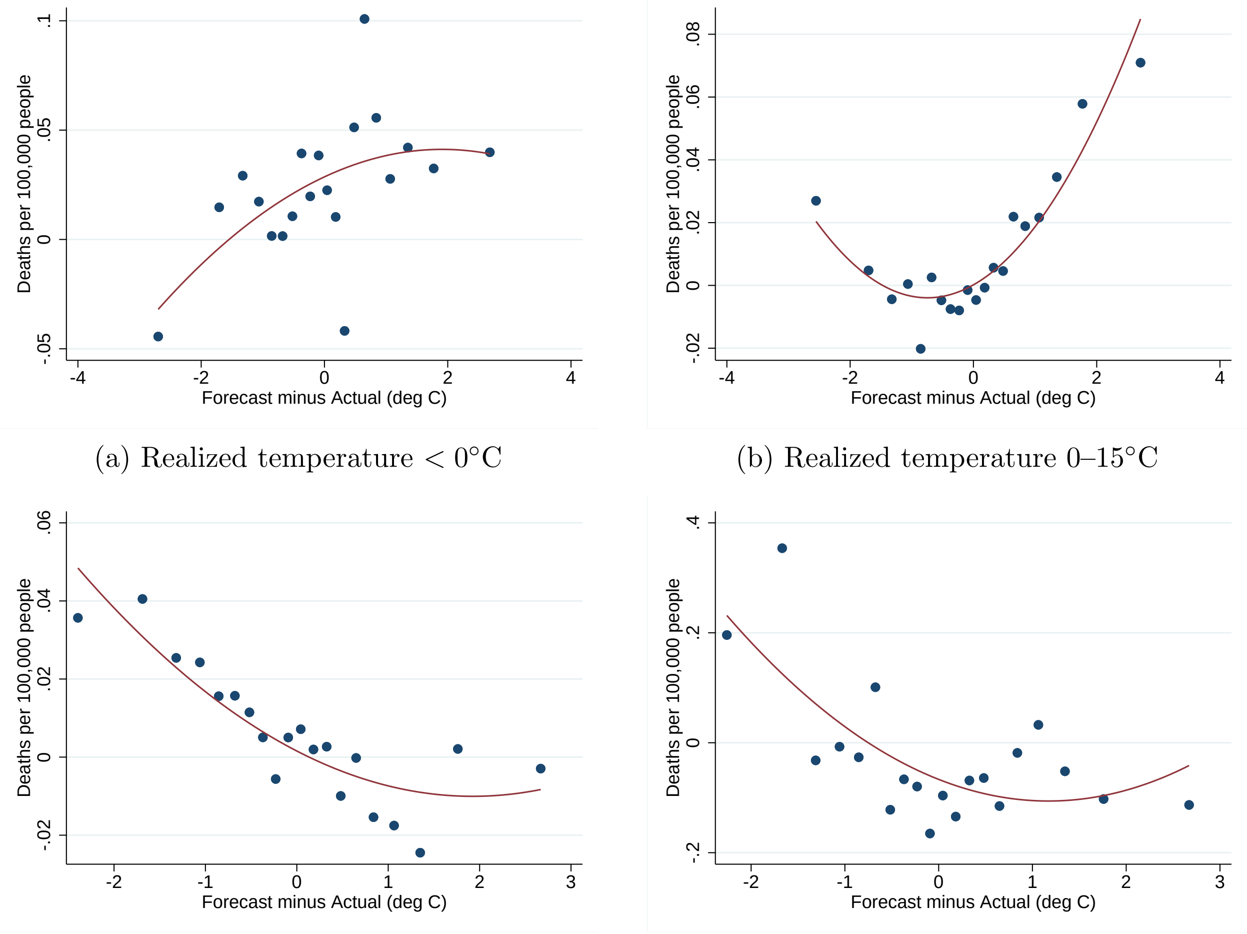
Notes: The maps show the standard deviation of the residuals from a regression of average temperature (left panel) or the 1-day-ahead forecast of average temperature (right panel) on all of the controls in the baseline regression specification (see Equation 10).

Figure A12: Spatial Variation in Forecast RMSE (Conditional on Baseline Controls)



Notes: The map shows the root mean squared error of the 1-day-ahead forecast for each county in the continental U.S. over the sample period. Redder values indicate higher average RMSE and yellower values indicate lower values. The values are all conditional on the baseline fixed effects and other control variables.

Figure A13: Raw Data Relationship Between 1-Day Ahead Forecast Error and Mortality for Days with Hot and Cold Realized Temperatures



(a) Realized temperature < 0°C

(b) Realized temperature 0–15°C

(c) Realized temperature 15–30°C

(d) Realized temperature > 30°C

Notes: The figures show bin-scatters of the relationship between forecast error from the 1-day ahead forecast and the daily mortality rate using the raw data residualized on realized temperature. The points are the average mortality rate within 20 quantiles of forecast error. Panel (a) shows the relationship when the expected temperature is cold (< 0°C), Panel (b) when cool (0–15°C), Panel (c) when warm (15–30°C), and Panel (d) when hot (> 30°C).

Table A4: Time Use: Quadratic Specification

	(1) Work	(2) Home prod.	(3) Leisure
$< 0°$ × Forecast error	4.01	-2.17	-1.84
	(4.20)	(4.18)	(6.95)
0 to 15° × Forecast error	0.35	-5.50***	5.16**
	(2.67)	(1.99)	(2.59)
15 to 30° × Forecast error	1.97	-1.01	-0.96
	(3.13)	(3.02)	(2.43)
$> 30°$ × Forecast error	-1.01	23.5*	-22.5*
	(10.2)	(13.9)	(11.8)
$< 0°$ × Forecast error^2	-5.30*	1.48	3.82
	(2.99)	(2.92)	(2.56)
0 to 15° × Forecast error^2	-1.06	1.13	-0.068
	(1.93)	(1.70)	(1.60)
15 to 30° × Forecast error^2	-5.46***	2.25	3.21*
	(1.91)	(1.71)	(1.88)
$> 30°$ × Forecast error^2	19.6**	6.42	-26.0***
	(8.99)	(7.07)	(9.59)
LHS mean	189.8	263.8	986.4
N	144,234	144,234	144,234
Clusters	100	100	100

Notes: The table shows estimates of the forecast error effect from a quadratic version of the time use regressions reported in Table 2. Additional covariates match the main results and are 5° bins for realized temperature, four bins for forecasted temperature, quartile bins for precipitation, as well as fixed effects for date, and location-by-month interacted with a linear time trend. Weighted by location population. Standard errors, clustered at the WFO level, are below each estimate. Significance: $p < .10$, ** $p < .05$, *** $p < .01$.

Table A5: Residential Electricity Demand: All Weather and Forecast Estimates

	(1)	(2)
	log electricity demand	
Temperature < −10°	0.013***	0.013***
	(0.0033)	(0.0034)
Temperature −10 to −5°	0.011***	0.011***
	(0.0026)	(0.0026)
Temperature −5 to 0°	0.011***	0.011***
	(0.0026)	(0.0026)
Temperature 0 to 5°	0.0077***	0.0077***
	(0.0015)	(0.0015)
Temperature 5 to 10°	0.0045***	0.0046***
	(0.0016)	(0.0016)
Temperature 10 to 15°	0.0032***	0.0032***
	(0.0012)	(0.0012)
Temperature 20 to 25°	0.0051***	0.0051***
	(0.0013)	(0.0013)
Temperature 25 to 30°	0.0087***	0.0087***
	(0.0021)	(0.0021)
Temperature > 30°	0.012***	0.012***
	(0.0025)	(0.0026)
Rain	-0.00029	-0.00021
	(0.00027)	(0.00027)
Rain × Rain	0.0000024	0.0000017
	(0.0000046)	(0.0000046)
Days < 0°	-0.00044	-0.00058
	(0.0013)	(0.0012)
Days 0 to 15°	-0.00010	-0.00020
	(0.00039)	(0.00040)
Days > 30°	-0.0010**	-0.0012**
	(0.00046)	(0.00048)
Days < 0° × Forecast error	0.00029	0.00023
	(0.00026)	(0.00025)
Days 0 to 15° × Forecast error	-0.00029	-0.00023
	(0.00020)	(0.00021)
Days 15 to 30° × Forecast error	0.00014	0.00016
	(0.00019)	(0.00019)
Days > 30° × Forecast error	0.0014***	0.0013***
	(0.00033)	(0.00033)
Baseline controls	Yes	Yes
Log price	No	Yes
Observations	7104	7104
Clusters	48	48

Notes: The table shows all estimates from the regression reported in Column 1 of Table 2 plus an additional version of the regression that includes the log of the electricity price. Standard errors, clustered at the state level, are below each estimate. Significance: $p < .10$, ** $p < .05$, *** $p < .01$.

Table A6: Residential Electricity Demand: Quadratic Specification

	(1)	(2)
	log electricity demand	
Days < 0° × Forecast error	0.000099	0.000030
	(0.00027)	(0.00026)
Days 0 to 15° × Forecast error	-0.00012	-0.000043
	(0.00024)	(0.00026)
Days 15 to 30° × Forecast error	0.0000047	0.000020
	(0.00018)	(0.00018)
Days > 30° × Forecast error	0.00094**	0.00093**
	(0.00038)	(0.00038)
Days < 0° × Forecast error^2	0.00038***	0.00039***
	(0.00010)	(0.00010)
Days 0 to 15° × Forecast error^2	-0.00033***	-0.00035***
	(0.000091)	(0.000088)
Days 15 to 30° × Forecast error^2	0.00033***	0.00034***
	(0.00012)	(0.00012)
Days > 30° × Forecast error^2	0.0011*	0.0011**
	(0.00055)	(0.00051)
Baseline controls	Yes	Yes
Log price	No	Yes
Cost ($)	3,397,047	3,383,431
	(1,712,338)	(1,853,717)
N	7104	7104
Clusters	48	48

Notes: The table shows estimates of the forecast error effect from a quadratic version of the residential electricity demand regression reported in Table 2. Additional covariates are 5° bins for realized temperature, four bins for forecasted temperature, a quadratic for precipitation, and the log of the price for residential electricity, as well as fixed effects for year-month, and state-by-month interacted with a linear time trend. Weighted by state population. Standard errors, clustered at the state level, are below each estimate. Significance: $p < .10$, ** $p < .05$, *** $p < .01$.

Appendix References

Barreca, A., K. Clay, O. Deschênes, M. Greenstone, and J. S. Shapiro (2016). Adapting to climate change: The remarkable decline in the U.S. temperature-mortality relationship over the 20th century. *Journal of Political Economy 124*(1), 105–159.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research.

CIESIN (2017). U.S. Census grids 2010. https://sedac.ciesin.columbia.edu/data/collection/gpw-v4. Accessed: 2018-08-30.

Deschênes, O. and E. Moretti (2009). Extreme weather events, mortality, and migration. *The Review of Economics and Statistics 91*(4), 659–681.

Gibson, M. and J. Shrader (2018). Time use and labor productivity: The returns to sleep. *Review of Economics and Statistics 100*(5), 783–798.

Heutel, G., N. H. Miller, and D. Molitor (2021). Adaptation and the mortality effects of temperature across us climate regions. *Review of Economics and Statistics 103*(4), 740–753.

Mesinger, F., G. DiMego, E. Kalnay, K. Mitchell, P. C. Shafran, W. Ebisuzaki, D. Jović, J. Woollen, E. Rogers, E. H. Berbery, et al. (2006). North american regional reanalysis. *Bulletin of the American Meteorological Society 87*(3), 343–360.

Missirian, A. (2020). Yes, in your backyard: Forced technological adoption and spatial externalities.

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

Simonsohn, U. (2018). Two lines: A valid alternative to the invalid testing of U-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science 1*(4), 538–555.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.