Estimating the Economic Impacts of Climate Stress on U.S. Dairy Production: Evidence from Debiased Machine Learning and Simulation-Based Estimator Choice

Duong Ta duongta@vt.edu

Mario Ortez omario@vt.edu

Bradford Mills bfmills@vt.edu

Michael Cary macary@vt.edu

Department of Agricultural and Applied Economics, Virginia Tech

Draft v1.5, last updated on October 19, 2026

Acknowledgments: The authors gratefully acknowledge the Department of Agricultural and Applied Economics at Virginia Tech for supporting this project by providing research facilities.

Funding: This research was funded by The Dairy Alliance under the grant titled "Enhancing dairy sector resilience through data-driven infrastructure planning and forecast modeling."

Conflict of Interest: The authors declare that they have no competing interests.

Abstract

We estimate how heat and cold exposure affect milk yield and fat content in a monthly panel of 21 major U.S. dairy states from 2003 through 2025. USDA NASS outcomes are matched to hourly ERA5-Land weather aggregated with county milk-cow inventory weights. State × calendar-month fixed effects absorb each state's seasonal cycle and year-month fixed effects absorb national monthly shocks, so responses are identified from differences in weather across states within the same month. We compare linear and cubic regressions with LASSO and neural-network learners, with and without automatic debiasing, and select the reported estimator for each stress tail based on its performance in synthetic-data simulations: the cubic regression for heat and the debiased neural network for cold, both with two-way clustered intervals. An additional 24-hour-equivalent day at or above THI 69 lowers milk per cow by 0.29 percent, cumulated over the exposure month and three lags. The cold estimate, −0.22 percent per day below THI 39, has a 95 percent interval that includes zero. Removing heat stress would have raised production by about 1.7 percent, and removing both tails by 2.9 percent. The size of these effects depends on the functional form of the response, which the data cannot pin down. Applied uniformly to 2003–2024 production in the 48 contiguous states, the joint share is worth about $23.9 billion, or $1.09 billion per year. Evaluated at each state's exposure, the heat share ranges from 0.7 percent of production in Oregon to 4.5 percent in Florida. Heat lowers and cold raises milk fat content. A break search finds no detectable change in the yield responses, but the heat effect on fat content weakens after the mid-2010s. Climate-normal diagnostics give suggestive but specification-sensitive evidence of adaptation.

1 Introduction

The U.S. dairy sector supports more than 3 million jobs and nearly $780 billion in annual economic activity (International Dairy Foods Association (IDFA), 2025). Dairy production also anchors rural demand for labor and services, while milk provides concentrated sources of calcium, protein, and vitamin D (Garcia & Kalscheur, 2004; Górska-Warsewicz et al., 2019; Heaney, 2014). This economic and nutritional role has been further amplified by growing consumer demand and commercial processing for protein-dense dairy products (e.g., OECD/FAO, 2022; USDA Economic Research Service, 2023), heightening the economic stakes of climate-driven yield and component shocks. How national milk production and fat content respond to the thermal environment determines the stability of this sector.

Temperature stress alters both milk volume per cow and milk composition. Heat suppresses feed intake and induces metabolic and endocrine responses that reduce mammary productivity (Kadzere et al., 2002; West, 2003). Heat can also lower fat and protein measures that matter for manufacturing yield and component-based payment (Moore et al., 2023; Vroege et al., 2023). Cold creates a different energy allocation problem: it raises the energy cows need for maintenance and thermoregulation, which can lower milk volume (Fu et al., 2022), even as measured fat content rises in winter (Heck et al., 2009; Salfer et al., 2019). A complete economic assessment must therefore examine both tails of the Temperature-Humidity Index (THI) distribution and keep quantity and composition on their distinct scales.

Identifying climate-stress responses is difficult even though short-run weather is outside producer control (exogenous to production decisions). Exposure follows seasonal patterns and differs across states alongside herd geography, productivity, and management. National shocks to feed costs, milk prices, and policy can coincide with weather that is correlated across states. We therefore first-difference the panel and include two sets of fixed effects (FE). State $\times$ calendar-month FE absorb each state's seasonal cycle and, because the panel is differenced, state-specific linear trends. Year-month FE absorb every national monthly shock. Responses are then identified from differences in weather across states within the same month. Because responses can be nonlinear and persist beyond the exposure month, we estimate the cumulative response over the exposure month and three lags.

We compare seven estimators. Linear OLS and Full OLS are linear and cubic series regressions; Full OLS year FE replaces the year-month FE with year FE. A LASSO with unpenalized time FE (UTF-LASSO) and a neural network (NNET) each enter as a plug-in estimator and with automatic debiasing (ADML), which adds a Riesz-representer correction (Chernozhukov et al., 2022; Klosin & Vilgalys, 2025). Rather than choosing an estimator after inspecting the estimates, we established decision rules and then ran a synthetic-data simulation that applies them. The simulation keeps the observed weather, generates synthetic outcomes from response shapes fitted to the data, and adds noise that preserves the observed cross-state correlation and persistence. An estimator qualifies if its 95 percent intervals contain the true effect in at least 93 percent of simulated samples (a coverage rate of at least 93 percent) under each of the three shapes used to evaluate the estimators. Among qualifying estimators, the rules select the one with the smallest worst-case bias, separately for heat and cold.

The rules select Full OLS for heat and the debiased neural network (NNET ADML) for cold, both with two-way clustered intervals. An additional 24-hour-equivalent heat day lowers milk per cow by 0.289 percent on the cumulative derivative scale (95 percent interval 0.110 to 0.468 percent). The debiased learners give somewhat larger heat estimates, in line with the overstatement the simulation documents for NNET ADML. The cold estimate is -0.220 percent per day, and its interval includes zero. The cold choice is also fragile: NNET ADML qualifies at exactly the coverage threshold, and every candidate estimator overstates the cold effect by about half under one plausible shape. Because five response shapes fit the data equally well and imply different effects, we report the range across them, a functional-form sensitivity range, alongside each estimate.

Four further results complete the picture. Heat lowers and cold raises milk fat content. A debiased valuation functional, which targets the production-weighted change in log yield directly, implies that removing heat stress would have raised production by about 1.7 percent and removing both tails, adding the two shares, by 2.9 percent. Applied uniformly to 2003–2024 production in the 48 contiguous states, that corresponds to $23.9 billion; evaluated at each state's own exposure, the heat share ranges from 0.7 percent in Oregon to 4.5 percent in Florida. Decade averages of the estimated scores do not differ detectably between 2006–

2015 and 2016–2025, and a break search finds no detectable change in the yield responses, although the heat effect on fat content weakens after the mid-2010s. Finally, climate-normal diagnostics suggest smaller heat damage where heat is historically common, but this evidence is sensitive to specification.

The paper contributes in three ways. It estimates both stress tails and both production margins in a common design that absorbs national monthly shocks. It selects among linear, regularized, and neural estimators by pre-specified simulation rules, and it shows how the debiased LASSO of Klosin and Vilgalys (2025) performs once the time FE are left unpenalized. It also values stress through a debiased functional and reports the range of effects across plausible response shapes rather than a single functional form. Section 2 reviews related work, Section 3 describes the data, Section 4 sets out the econometric framework, and Section 5 presents the simulation. Sections 6–8 report the estimates, the temporal and adaptation diagnostics, and the economic magnitudes, and Section 9 concludes.

2 Related literature and contribution

2.1 Animal-science dose response

Animal-science research provides the physiological mechanism and threshold ranges that motivate our exposure design. Dairy cows generate substantial metabolic heat through rumen fermentation and lactation. When heat dissipation is constrained, respiration and body temperature rise, feed intake falls, and milk synthesis declines (Armstrong, 1994; Kadzere et al., 2002; West, 2003). Historical work often used THI 72 as the onset of heat stress, but evidence for modern high-producing cows places the threshold closer to 68 or 69 (McWhorter et al., 2023). We therefore define heat exposure at THI 69 rather than waiting for more severe conditions.

Cold stress receives less attention, partly because lactating cows generate substantial internal heat. Yet the lower tail matters in large observational panels. U.S. evidence shows productivity losses below approximately THI 39 and a gradual cold-response gradient rather than a single sharp break (Choi et al., 2024; Gisbert-Queral et al., 2021). We use THI 39 as the lower boundary of the reference range. These thresholds summarize exposure rather than claim that every breed, housing system, or lactation stage shares an identical biological

cutoff.

2.2 Measuring climate stress in agricultural economics

Agricultural climate economics represents nonlinear thermal exposure in two related ways. Duration measures record time spent in temperature bins or beyond a threshold: Schlenker and Roberts (2009) reconstruct the within-day temperature distribution, Zaveri and Lobell (2019) count growing-season days above 30 degrees Celsius, and Tack et al. (2015) combine duration below freezing with heat degree days. Such measures preserve damaging tails that seasonal averages obscure (Hsiang, 2016). Degree-day measures instead weight exposure by its distance beyond a threshold; their use in crop-yield, insurance, and adaptation studies has made them a standard convention (Annan & Schlenker, 2015; Burke & Emerick, 2016; Schlenker & Roberts, 2009; Tack et al., 2015).

Dairy research draws the same distinction with THI. St-Pierre et al. (2003) calculate both heat-stress duration and `THI_Load`, the integral of THI above a threshold, and Mbuthia et al. (2022) show that a similar hourly accumulation, "cooling THI-days," adds information beyond conventional THI models. Our `THI69_Days` measure records duration: every hour at or above THI 69 contributes 1/24 of a stress day, irrespective of severity. `THI69_Load` is the intensity-duration counterpart, closer to degree days. Section 3.3 defines both and Section 6.4 reports estimates for both; their coefficients are not directly comparable because their units differ.

2.3 Damage estimates

Heat stress imposes well-documented costs on U.S. dairy production. St-Pierre et al. (2003) combine animal-response functions with long weather histories and estimate average annual losses near $897 million without heat abatement. With operation-level survey data and finely scaled climate records, Key et al. (2014) find that heat stress reduced the average farm's 2010 production value by roughly $39,000, with losses concentrated in southern states. Mauger et al. (2015) estimate historical national yield losses of 1.9 percent relative to an idealized no-heat-stress baseline (1950–1999 climate, 2009 cow populations, no heat abatement), rising to 6.3 percent by the end of the century. Cooling can offset part of these losses, though the response varies across climate zones and production systems (Lopez et al.,

2022; Mukherjee et al., 2013), and Hutchins et al. (2025) find the greatest heat vulnerability in small herds, suggesting barriers to adaptation.

Heat also degrades milk composition, which matters for revenue under component pricing. Ravagnolo et al. (2000) show that fat and protein yields decline once THI exceeds 72. Vroege et al. (2023) find that composition losses amplify the revenue effect of lower volume, and Hutchins et al. (2025) that quality effects on farm income exceed the volume effect. Most directly relevant to our setting, Prieto et al. (2026) link 2007–2016 test-day records for roughly 6.5 million U.S. cows to gridded THI and show that fat and protein content begin declining at lower THI than yield. Priced through Federal Milk Marketing Order component rates, their revenue damage under a uniform ten-unit THI increase is about twice the yield-only estimate, 2.8 percent or about $1.65 billion at 2022 national milk revenue, in a warming scenario that differs from our observed-exposure accounting.

At our geographic unit, Fan and Ma (2024) use spatial-error and spatial-Durbin models of a U.S. state-year panel and find an inverted-U relationship between heat stress and milk per cow. They show that two-way fixed-effects residuals carry spatial dependence, which matters for inference and possibly identification. We do not model spatial errors; our intervals instead allow arbitrary correlation across states within a month and within states over time.

Cold stress has received less attention in economic studies. In monthly U.S. state data for 1981–2018, Gisbert-Queral et al. (2021) associate extreme heat above THI 79 and cold below THI 39 with daily yield reductions of about 3.7 and 6.1 percent relative to THI 65–69, and Choi et al. (2024) find continuous nonlinear responses in both tails across four thermal indices for 1981–2020. Seasonal studies report higher winter milk fat concentration (e.g., Heck et al., 2009; Salfer et al., 2019), but seasonal patterns alone do not identify a cold-stress effect, and cold-related composition responses remain little studied. These findings motivate our two-sided design.

Evidence on response change differs. Choi et al. (2024) report lower yield sensitivity in the most recent two decades of their panel, whereas Prieto et al. (2026) find similar heat-response curves for 2007–2011 and 2012–2016 and caution that outcome stability does not establish the presence or absence of adaptation, because management practices are unobserved.

3 Data and exposure construction

3.1 Outcomes

Monthly milk production per cow and milk fat content come from USDA NASS QuickStats (USDA National Agricultural Statistics Service (NASS), n.d.). The balanced production panel covers 21 major dairy states from 2003 through 2025 and represents roughly 90 percent of U.S. milk output. The quantity outcome is log monthly pounds of milk per cow (`lmpc`). The composition outcome is milk fat content (`fc`) in percentage points.

Both margins matter economically. Volume per cow determines raw milk supply and farm receipts, while fat content determines manufacturing yields for products such as butter and cheese and payments under Federal Milk Marketing Order (FMMO) component pricing. Both are co-primary outcomes in the estimation. The monetary accounting covers quantity only, because valuing fat content would require a component-pricing framework that combines composition with quantity. The temporal diagnostics cover both outcomes, except the climate-normal diagnostics, which use `lmpc`.

USDA records omit milk per cow for March through June 2013. We impute these months with the state-calendar-month average of 2012 and 2014, which keeps a complete calendar for first differences and lags. Treating them as missing instead changes the Full OLS cumulative effects only slightly, from -0.289 to -0.292 percent per day for heat and from -0.237 to -0.228 percent for cold. Fat content is missing for Colorado in 2003–2004 and for Kansas and Oregon in 2004 (48 observations), and we leave it missing rather than impute composition. Lags are built on the complete monthly calendar, and differences use only adjacent observed months, so no difference spans a missing month.

3.2 Hourly weather and THI

ERA5-Land supplies hourly 2-meter air temperature, dew point, and precipitation. We convert temperature and dew point from Kelvin to degrees Celsius and calculate relative humidity RH using the standard Magnus approximation, and then apply the Yeck THI formula:

$$THI = 1.8T + 32 - (0.55 - 0.0055RH)(1.8T - 26),$$

where T is 2-meter air temperature in degrees Celsius and RH is relative humidity in percent.

THI is computed hourly for each grid cell, accumulated into 24-hour-equivalent exposure days within each county-month, and aggregated from counties to states with milk-cow inventory weights (Section 3.4). Monthly precipitation is the sum of hourly millimeters.

3.3 Climate stress measurements

The main model uses duration-day measures. For a county-month with H hourly observations, we define

$$\texttt{THI39_Days} = \frac{1}{24}\sum_{h=1}^{H}\mathbf{1}(THI_h < 39), \qquad \texttt{THI69_Days} = \frac{1}{24}\sum_{h=1}^{H}\mathbf{1}(THI_h \geq 69).$$

Each stress day represents 24 possibly nonconsecutive stressful hours within the county-month, without extra weight for more severe hours. The omitted thermoneutral range is $39 \leq \text{THI} < 69$. The thresholds follow the lower-tail U.S. evidence and the lower heat threshold observed for modern high-producing cows (Choi et al., 2024; Gisbert-Queral et al., 2021; McWhorter et al., 2023).

The model adds quadratic and cubic terms in both measures and their interactions with precipitation, but duration terms do not capture thermal intensity. As an intensity-weighted alternative, load measures accumulate each hourly deviation beyond the same thresholds:

$$\texttt{THI39_Load} = \sum_{h=1}^{H}\max(0, 39 - THI_h), \qquad \texttt{THI69_Load} = \sum_{h=1}^{H}\max(0, THI_h - 69).$$

Load is measured in THI-hours per county-month, so an hour farther beyond the threshold contributes more. Because raw loads are highly skewed and can be zero, estimation uses shifted logs:

$$\texttt{ln_THI39_Load} = \ln(0.001 + \texttt{THI39_Load}),$$
$$\texttt{ln_THI69_Load} = \ln(0.001 + \texttt{THI69_Load}).$$

The 0.001 shift retains zero-load months, and the logarithm compresses the upper tail. For log yield, the derivative with respect to shifted-log load approximates a raw-load elasticity where load is large relative to 0.001; for fat content, its units are percentage points per shifted-log-load unit. The two-bin design trades resolution for statistical power in a 21-state panel and recovers within-tail nonlinearity through the cubic and interaction terms.

3.4 Dairy-cow-weighted spatial aggregation

Averaging weather over land without dairy cattle can attenuate extremes. Crop studies address this aggregation bias by weighting exposure toward cropland or planted area (Burke & Emerick, 2016; Deschênes & Greenstone, 2007; Schlenker & Roberts, 2009); we weight county exposure by time-varying milk-cow inventories, so the state-month treatment measures weather where cows are located. This matters most in large states with concentrated dairy regions.

We first aggregate ERA5-Land grid cells to counties. County exposures are then averaged within each state-month using county milk-cow inventories from the 2007, 2012, 2017, and 2022 agricultural censuses. Each sample year uses its nearest designated census anchor: 2007 for 2003-2009, 2012 for 2010-2014, 2017 for 2015-2019, and 2022 for 2020-2025. Counties without supported cow inventory receive zero weight.

3.5 Descriptive statistics

3.5.1 Summary statistics

Table 1 summarizes the main variables before differencing and lag construction. After lags and first differences, the yield sample has 5,712 state-months (May 2003 through December 2025) and the fat sample 5,665, because the first difference after each fat-content gap is dropped.

Table 1. Summary statistics for the main state-month modeling variables.

Variable	Unit	N	Mean	SD	Min	P25	Median	P75	Max
`lmpc`	log lb/cow	5,796	7.500	0.122	6.928	7.415	7.507	7.596	7.783
`mpc`	lb/cow	5,796	1,821	218	1,020	1,660	1,820	1,990	2,400
`fc`	percent	5,748	3.824	0.241	3.310	3.650	3.780	3.950	4.760
`THI69_Days`	days/mo	5,796	6.223	8.473	0.000	0.000	1.345	10.613	31.000
`THI39_Days`	days/mo	5,796	6.601	9.191	0.000	0.000	1.004	11.609	30.997
`precip_mm`	mm/mo	5,796	71.682	49.265	0.023	32.832	65.613	100.884	373.098

Note: `lmpc` is log milk production per cow; `mpc` is monthly milk production per cow in pounds; `fc` is milk fat content in percent; `THI69_Days` is days with THI ≥ 69; `THI39_Days` is days with THI < 39; `precip_mm` is monthly precipitation in millimeters. Sample covers 21 major dairy states from 2003 through 2025.

3.5.2 Temporal and spatial patterns

Figure 1 shows the monthly dynamics across the 21 study states. Heat exposure peaks in July and August and cold exposure in December through February (Panels A and B). Milk per cow (Panel C) combines a strong seasonal cycle, peaking in spring and falling in summer, with a long-run rise; fat content (Panel D) falls in summer, rises in cold months, and trends upward in later years. These aggregate series do not separate the roles of genetics, nutrition, technology, or pricing incentives.

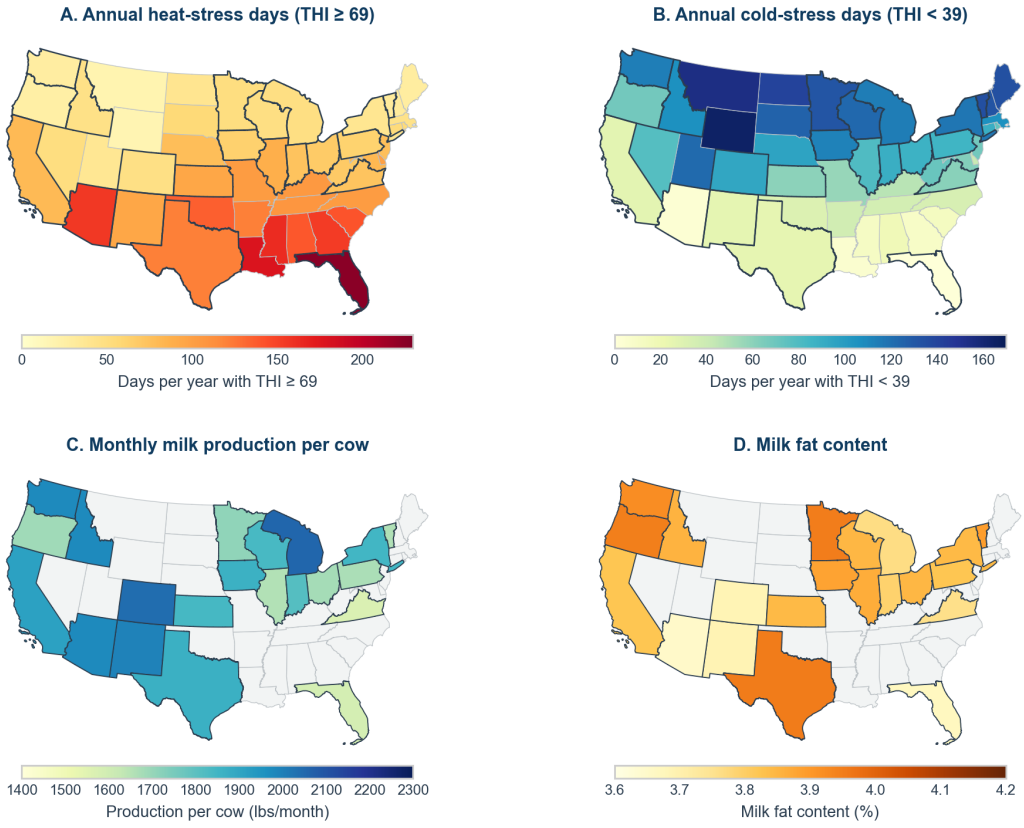
Figure 1. Monthly temporal variations in weather stress exposure and dairy productivity (2003-2025).



Note: Panel A plots the monthly average heat-stress days ($THI69_Days$, THI $\geq$ 69); Panel B plots the monthly average cold-stress days ($THI39_Days$, THI $<$ 39); Panel C plots monthly average milk production per cow (mpc in pounds); Panel D plots monthly average milk fat content (fc in percent). All series are equally weighted state averages over 2003–2025. Panels A–C use all 21 study states. Panel D uses available fat-content observations: 20 states in 2003, 18 in 2004, and 21 thereafter.

Figure 2 maps long-run averages. Heat exposure is concentrated in southern and southwestern states such as Texas, Florida, and Arizona, where herds experience over 100 heat-stress days per year, and cold exposure in northern and midwestern states such as Wisconsin, Minnesota, and Idaho. Yield per cow is highest in the West and Southwest (e.g., Colorado, California, Idaho) and fat content in northern states (e.g., Oregon, Minnesota, Vermont), although the maps do not separate climate from herd composition or management.

Figure 2. Spatial distribution of average annual weather stress and dairy productivity (2003-2025).



Note: Panel A maps average annual heat-stress days (THI $\geq$ 69); Panel B maps average annual cold-stress days (THI $<$ 39); Panel C maps long-run average monthly milk production per cow (lbs/month); Panel D maps long-run average milk fat content (percent). Weather panels cover the 48 contiguous states. The 21 study states are highlighted with dark outlines. Production and fat averages use available observations; non-study states in Panels C and D are light grey.

4 Econometric framework

4.1 Targets and identifying assumptions

Let W_{it} collect the heat and cold exposures and precipitation of state i in month t and the three preceding months. The level model is

$$y_{it} = \mu_{i,m(t)} + \kappa_i t + \lambda_t + g_0(W_{it}) + \varepsilon_{it}, \qquad \mathbb{E}[\varepsilon_{it} \mid \mu, \kappa, \lambda, \{W_{is}\}_s] = 0. \tag{1}$$

Here y_{it} is log monthly milk production per cow (`lmpc`) or milk fat content in percent (`fc`). The state $\times$ calendar-month FE $\mu_{i,m(t)}$ absorb each state's seasonal cycle, the seasonal

confounding that arises because stress exposure and production follow annual cycles that differ across states. $\kappa_i t$ is a state-specific linear trend, and the year-month FE λ_t absorb any national shock in month t. The response function g_0 is common to all states, a maintained restriction on state-level averages rather than on individual herds. Strict exogeneity excludes omitted time-varying determinants that differ across states within a month and are correlated with the exposure history. Weather being outside producer control supports this condition but does not establish it. In particular, the year-month FE do not remove regional shocks that coincide with regional weather.

For stress tail $r \in \{C, H\}$ with exposure D^r, the per-day effect is the cumulative average partial derivative over lags 0–3:

$$\delta_r = \sum_{\ell=0}^{3} \mathbb{E}\left[\frac{\partial g_0(W_{it})}{\partial D^r_{i,t-\ell}}\right]. \tag{2}$$

For duration measures, an additional 24-hour-equivalent stress day replaces thermoneutral exposure ($39 \leq \text{THI} < 69$) at the same lag, holding the other tail and precipitation fixed. For `lmpc`, $100\delta_r$ is the approximate percentage response on the cumulative derivative scale; it is not a percentage loss in total milk over four months. For `fc`, δ_r is in percentage points. Because the year-month FE absorb every national monthly change, δ_r is identified from differences in exposure across states within a month.

The second target is a production share. For a counterfactual W^r_{it} that removes stress tail r in the current and three preceding months,

$$\theta_r = \frac{\sum_{i,t} w_{it}\{g_0(W^r_{it}) - g_0(W_{it})\}}{\sum_{i,t} w_{it}}, \tag{3}$$

where w_{it} is the state's milk production in month t. Removing a duration tail sets its stress days to zero; removing a load tail sets the shifted-log load to its zero-load value, $\ln(0.001)$. The joint counterfactual W^{CH}_{it} removes both tails and defines θ_{CH} in the same way. For `lmpc`, θ_r is the production-weighted mean change in log yield. The proportional gain in production, $\sum_{i,t} w_{it}\{e^{d_{it}} - 1\} / \sum_{i,t} w_{it}$ with $d_{it} = g_0(W^r_{it}) - g_0(W_{it})$, is at least $e^{\theta_r} - 1$ and differs from it because d_{it} varies across state-months; for the Full OLS stress-day shares the

difference is under 1 percent of the gain (Appendix F). Unlike the product of an average derivative and average exposure, θ_r evaluates the nonlinear response at each state-month's observed exposure.

4.2 Basis, transformation, and estimators

The weather basis $b(W_{it})$ contains, at each of the four lags, linear, quadratic, and cubic terms in both stress measures and precipitation, plus the two stress-by-precipitation interactions: 44 weather columns. Lags are built on the complete monthly calendar. The estimating equation first-differences (1) and then removes the mean of each differenced variable within state $\times$ calendar month, using only that state's rows. This removes $\mu_{i,m(t)}$ and κ_i exactly. The year-month FE enter as 272 indicator columns that are differenced and demeaned along with the weather basis and are never penalized. Differencing requires the outcome in two adjacent months, so the yield sample has 5,712 state-months (May 2003 through December 2025) and the fat sample has 5,665.

The seven estimators share this transformed design.

- **Linear OLS** uses the linear weather terms and the FE. **Full OLS** uses the full cubic basis; its per-day effect averages the analytical derivative of the fitted basis. **Full OLS year FE** replaces the year-month FE with year FE and is reported as a comparison.
- **UTF-LASSO Plug-in** fits a LASSO to the weather columns after partialling the FE out of the outcome and the weather columns, so the FE are unpenalized; its per-day effect averages the derivative of the fitted function. **NNET Plug-in** uses a feed-forward network f that maps the level weather history to a level response. Its differenced prediction $f(x_{it}) - f(x_{i,t-1})$, plus linear terms in the transformed FE, is transformed like the outcome and fitted by least squares. The network's derivative $\partial f(x_{it})/\partial D^r_{i,t-\ell}$ is therefore a level derivative by construction. Weight decay applies to the network weights but not to the FE.
- **UTF-LASSO ADML** and **NNET ADML** add the Riesz correction described below to the two plug-ins (Klosin & Vilgalys, 2025).

For the debiased estimators, the held-out score for lag ℓ is

$$\widehat{\psi}_{r\ell,it} = \hat{q}_{r\ell,it} + \hat{\alpha}_{r\ell}(Z_{it})\{\Delta\tilde{y}_{it} - \hat{f}(Z_{it})\}, \qquad \hat{\delta}_r = \frac{1}{N}\sum_{i,t}\sum_{\ell=0}^{3}\widehat{\psi}_{r\ell,it}, \tag{4}$$

where $\hat{q}$ is the plug-in derivative, $\Delta\tilde{y}$ the transformed outcome, $\hat{f}$ the fitted transformed prediction, and $\hat{\alpha}_{r\ell}(Z) = Z'\hat{\rho}_{r\ell}$ a linear Riesz representer (Chernozhukov et al., 2022). The representer is fitted on the transformed basis Z, with the FE partialled out so that they are not penalized. For NNET ADML this basis also includes natural cubic spline terms in each exposure and lag. Their knots are the exposure floor, the quartiles of exposure above it, and the maximum, which lets the representer bend where the cubic dictionary cannot. The ℓ_1 penalty is the theory rate times a multiplier chosen by held-out-state Riesz loss (Appendix A). The production share (3) uses the same scores with the derivative direction replaced by the production-weighted counterfactual change $w_{it}\{b(W_{it}^r) - b(W_{it})\}$. For Full OLS, the score of (2) or (3) is the influence function of a linear combination of the OLS coefficients, so Full OLS uses the same inference formulas.

4.3 Cross-fitting and inference

The learners use five outer folds that hold out whole states, so no observation from an evaluation state enters its nuisance fit. To remove dependence on a single fold assignment, we repeat cross-fitting over 20 assignments and report the median estimate. Each standard error is $\{\text{median}_s(\widehat{se}_s^2 + (\hat{\delta}_s - \tilde{\delta})^2)\}^{1/2}$ over assignments s, with $\tilde{\delta}$ the median (Chernozhukov et al., 2018). OLS estimators are fitted once on the full sample.

Holding out states prevents overlap between the training and evaluation samples but not dependence between them, because residuals of different states in the same month can be correlated. The cross-fitting therefore assumes that the dependence left after the year-month FE is weak enough for the nuisance fits to be approximately independent of the held-out residuals. The simulation in Section 5 supports this for the per-day effect: its noise keeps the observed cross-state correlation within each month, and the cross-fitted estimators still reach the coverage reported there. Holding out months as well, as multiway cross-fitting does (Chiang et al., 2022), would leave the year-month FE of held-out months without training data and would cut through the lag structure, so we do not use it.

Residual correlation across states within a month is the main inferential concern in a state panel, because weather and its consequences are spatially correlated (Fan & Ma, 2024). Year-month FE remove the national component of this correlation, but regional correlation within a month remains. We report two variance estimators of the score mean.

- The two-way clustered variance clusters by state and by sample month and allows arbitrary correlation within either dimension (Cameron et al., 2011). Intervals use $t(20)$ critical values for 21 states.
- The Driscoll–Kraay variance sums scores across states within each month and applies a Bartlett kernel with a 12-month bandwidth to the monthly sums. It allows cross-sectional dependence of unknown form and serial correlation (Driscoll & Kraay, 1998). Intervals use normal critical values.

State-clustered standard errors are shown for comparison only. The simulation in Section 5 shows that they understate uncertainty in this design. Following decision rule 1 of Section 5, the text quotes two-way intervals, and every table also shows the Driscoll–Kraay standard errors.

4.4 Lag window

We choose the lag window by adding lags one at a time on a fixed sample, from the exposure month alone (0–0) to twelve lags (0–12). For each estimator and stress tail, the window stops before the first newly added lag whose two-way 95 percent interval includes zero. The rule keeps lags 0–3 if every stopping point is at lag 3 or earlier, and it never shortens the window below 0–3. The procedure was run for Linear OLS, Full OLS, UTF-LASSO ADML, and NNET ADML for both outcomes. Every two-way stopping point is at lag 3 or earlier, so all estimates use lags 0–3. Appendix D reports the stopping points. Stopping points beyond lag 3 arise only under the Driscoll–Kraay and state-clustered intervals, for the NNET ADML fat-heat response.

An insignificant added lag does not show that later lags have no effect, so Appendix Table A4 (Panel C) also reports the selected estimators' cumulative effects over windows 0–3 to 0–6 on the same sample, with the sum of the lag 4–6 effects within the 0–6 window. The cumulative yield effects grow with the window, from −0.283 to −0.431 percent per day

for heat (Full OLS) and from −0.201 to −0.515 percent for cold (NNET ADML, one fold assignment), and they become less precise. The lag 4–6 effects are not significant for either tail or outcome (two-way p-values 0.16 to 0.67). For heat they are small (−0.040 percent); most of the change comes from the lag 0–3 terms, which shift when later lags enter. We keep lags 0–3. The production shares remove stress at these lags only, so they would understate the loss if later lags add damage.

5 Simulation evidence and estimator choice

5.1 Design

The simulation asks which estimators recover the per-day effect and give reliable intervals in this panel, and how much the effect depends on the shape of the response. It is a synthetic-data simulation: it keeps the observed weather and the 5,712-row yield sample fixed and generates synthetic outcomes as

$$y_{it}^{\text{sim}} = \text{base}_{it} + g_T(W_{it}) + u_{it}. \tag{5}$$

The baseline and the noise come from a levels regression of log yield on state × calendar-month indicators, state-specific linear trends, year indicators, and the weather basis. The baseline is the fitted non-weather component. The noise u_{it} reuses the regression residuals: each replication applies one circular time shift, one sign, and one optional time reversal to every state's residual series. Because the same transformation applies to all states, the noise keeps the observed correlation across states within a month, including common national shocks, and each state's persistence.

The structural component g_T takes one of five response shapes, each fitted to the observed data in the estimation space (year-month FE, lags 0–3): linear; cubic, which equals the Full OLS fit; convex, built from hinge terms so that damage accelerates with exposure; concave, built from capped terms so that damage saturates; and a natural cubic spline. The convex and concave fits constrain the stress terms to be monotone in exposure; their interactions with precipitation are left free. Each shape is zero at zero exposure and shares the precipitation terms. The true per-day effect (2) and production share (3) of every shape are exact functions of its fitted coefficients.

The design has two tiers. The first tests whether the data can distinguish the shapes. It runs 500 replications of Linear and Full OLS under each shape and compares two fingerprints with their observed values: the difference between the Linear and Full OLS heat effects, and the same difference for cold. A shape passes if each observed fingerprint lies within the central 95 percent of its simulated distribution. All five shapes pass (Appendix Table A3): the observed fingerprints fall at the 28th to 79th percentiles for heat and the 48th to 85th for cold, and the fits' R^2 differ by at most 0.012. The data therefore cannot tell these shapes apart.

The second tier runs all seven estimators, 500 replications under each of three shapes. Under the cubic shape, Full OLS is correctly specified, so its coverage measures standard-error calibration alone. The natural-spline shape departs from the cubic dictionary shared by the OLS and LASSO arms and by the Riesz representers. Under the convex shape, damage accelerates with exposure. The linear and concave shapes enter the first tier and the sensitivity range but are not replicated with the learners.

5.2 Decision rules

The following rules were established before the simulation was run.

1. **Inference.** An estimator qualifies if its two-way or its Driscoll–Kraay interval covers the true effect at least 93 percent of the time under every second-tier shape. Coverage above 97 percent counts as conservative rather than as failure. The text quotes the two-way interval if it qualifies, and the Driscoll–Kraay interval otherwise.
2. **Estimator.** The candidates are Linear OLS, Full OLS, UTF-LASSO ADML, and NNET ADML, all with year-month FE. Separately for heat and cold, the reported estimator is the qualifying candidate with the smallest worst-case absolute relative bias across the second-tier shapes. Candidates within two paired Monte Carlo standard errors of that minimum are tied, and a tie goes to the smaller worst-case root mean squared error (RMSE). The plug-ins and Full OLS year FE are comparisons, not candidates.
3. **Headline target.** If the true per-day effect differs by more than 25 percent across the shapes that pass the first tier while the production share differs by less, the paper

leads with the production share. If both differ by more than 25 percent, the paper reports a range across plausible shapes.

5.3 Results

Table 2 reports relative bias and coverage, and Figure 3 shows the error distributions. For heat, three candidates qualify with two-way intervals: Full OLS covers 94.6 to 96.0 percent of the time, UTF-LASSO ADML 96.6 to 97.4 percent, and NNET ADML 96.8 to 97.4 percent. Linear OLS fails, covering only 81.4 to 91.6 percent. Full OLS has the smallest worst-case bias, a 9.9 percent overstatement of the damage under the natural-spline shape. UTF-LASSO ADML is tied with it (10.4 percent; the paired difference is 0.6 percentage points with a paired Monte Carlo standard error of 0.7). The tie goes to Full OLS, whose worst-case RMSE is lower (0.98 against 1.19 $\times 10^{-3}$ log points). NNET ADML overstates heat damage under every shape, by 8.7 to 14.3 percent, and is not tied. Its plug-in recovers only about a third of the effect, so the correction carries the estimate and passes on the approximation error of its Riesz representer.

For cold, the shapes disagree. Under the cubic and convex shapes Full OLS and both debiased learners are within 7 percent of the truth. Under the natural-spline shape every candidate overstates the cold effect by roughly half: by 54.9 percent for Full OLS, 57.5 percent for UTF-LASSO ADML, and 46.4 percent for NNET ADML. This bias lowers the two-way coverage of Full OLS and UTF-LASSO ADML to 81.6 and 86.4 percent under that shape, so neither qualifies. NNET ADML qualifies because its bias is somewhat smaller and its spread larger: its lowest coverage is 93.0 percent, exactly the threshold. The rule therefore selects NNET ADML for cold, but with one more miss in 500 replications no estimator would have qualified. We report the cold estimate with this fragility in view.

The inference results support the choice of intervals. For Full OLS and the two debiased learners, the mean two-way standard error is 0.93 to 1.07 times the standard deviation of the estimates. The Driscoll–Kraay standard error is 4 to 11 percent too small, and the state-clustered standard error is 10 to 24 percent too small. State clustering therefore understates uncertainty even after the year-month FE remove the national shock, because residuals remain correlated across states within a month. The plug-ins are badly biased: UTF-LASSO Plug-in covers 12 to 29 percent of the time, and NNET Plug-in understates

the heat effect by 60 to 70 percent and never covers. Full OLS year FE has bias similar to Full OLS, but its standard deviation is 16 percent (heat) and 26 percent (cold) larger, so year-month FE also improve precision.

The exact true effects of the five shapes span a wide range. The per-day heat effect runs from -0.219 to -0.363 percent, a 66 percent spread, and the cold effect from -0.146 to -0.237 percent (63 percent). The production share removed by heat runs from 1.28 to 2.12 percent (66 percent), by cold from 0.98 to 1.37 percent (40 percent), and by both from 2.26 to 3.50 percent (55 percent). Under rule 3 we therefore report each estimate with the range across the plausible shapes. This is a functional-form sensitivity range: it shows how much the effect depends on which shape holds, not an interval for the true effect, and the shapes' fitted coefficients carry sampling error of their own. The simulation evaluates the per-day effect on milk yield with the stress-day measures, fitting each learner on one fold assignment per replication. It does not evaluate the median over 20 assignments, which reduces the variation from the fold assignment and widens the standard error (Section 4.3), nor the production shares or their sum. Fat content and the load measures are not simulated.

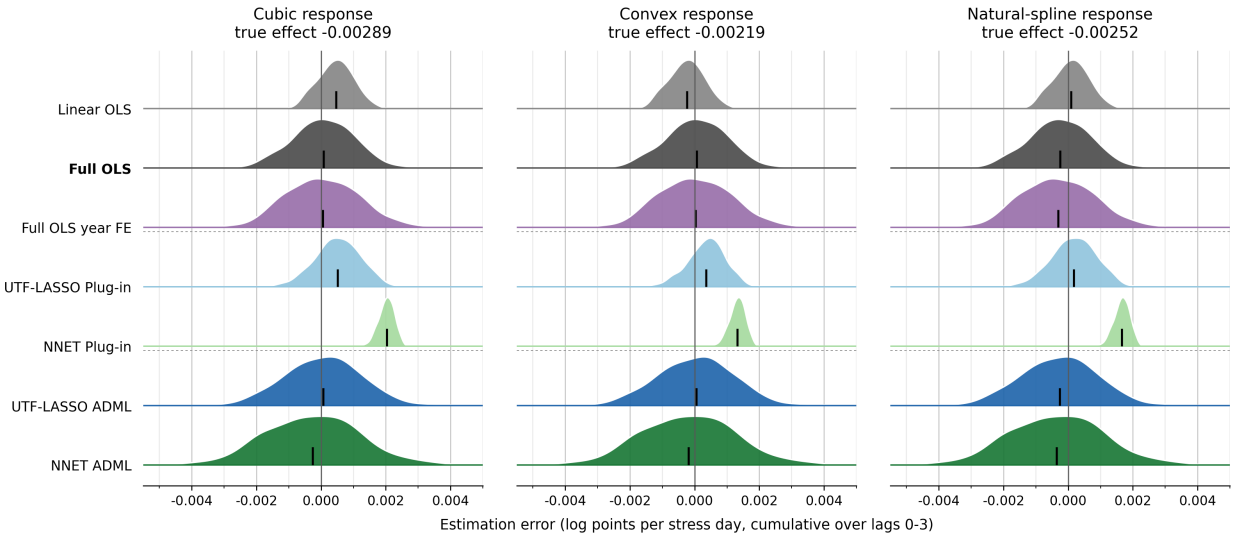
Table 2. Simulation results for the cumulative per-day effect: relative bias and interval coverage.

	Relative bias (%)				Coverage (range)	
Estimator	Cubic	Convex	Spline	SD	Two-way	DK
Panel A. Heat (THI $\geq$ 69)						
Linear OLS	+15.8	−11.1	+3.6	0.55	0.814–0.916	0.728–0.854
Full OLS	+2.6	+3.1	−9.9	0.95	**0.946–0.960**	0.906–0.928
Full OLS year FE	+1.6	+1.1	−12.2	1.11	0.956–0.968	0.926–0.946
UTF-LASSO Plug-in	+17.6	+16.4	+7.0	0.65	0.174–0.286	0.022–0.026
NNET Plug-in	+70.1	+60.5	+65.9	0.24	0.000–0.000	0.000–0.000
UTF-LASSO ADML	+2.4	+2.1	−10.4	1.16	**0.966–0.974**	0.912–0.928
NNET ADML	−9.2	−8.7	−14.3	1.43	**0.968–0.974**	**0.936–0.942**
Panel B. Cold (THI $<$ 39)						
Linear OLS	+17.4	−36.0	−25.3	0.36	0.684–0.806	0.638–0.750
Full OLS	−1.0	−6.8	−54.9	0.88	0.816–0.944	0.780–0.926
Full OLS year FE	−1.8	−8.0	−51.8	1.11	0.900–0.940	0.862–0.892
UTF-LASSO Plug-in	+28.6	+9.2	−8.2	0.48	0.118–0.280	0.012–0.050
NNET Plug-in	+68.2	+47.7	+51.9	0.18	0.000–0.000	0.000–0.000
UTF-LASSO ADML	−2.1	−4.8	−57.5	1.02	0.864–0.942	0.818–0.922
NNET ADML	−5.2	−2.2	−46.4	1.21	**0.930–0.964**	0.882–0.922

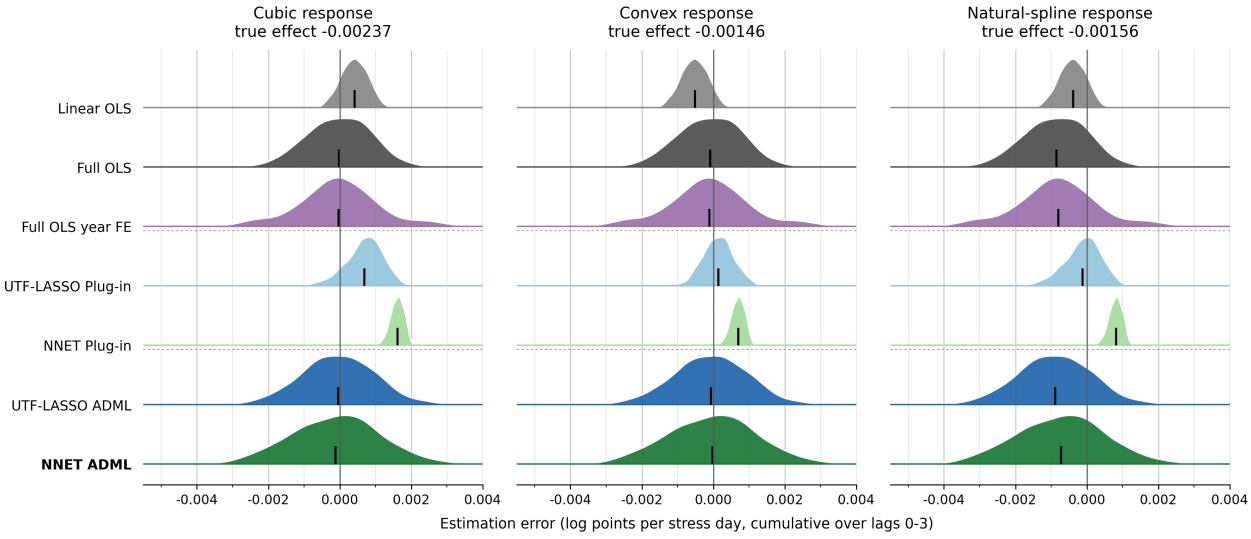
Note: 500 replications per shape; yield, stress days, lags 0–3. Relative bias is mean error over the absolute true effect. The true effects are negative, so a negative bias overstates the damage. SD is the standard deviation of the estimates in 10^{-3} log points per stress day, averaged over the three shapes. Coverage is the lowest-to-highest share of nominal 95 percent intervals containing the true effect across the three shapes. Two-way intervals use $t(20)$ critical values; Driscoll–Kraay (DK) intervals use normal critical values. Bold coverage ranges qualify under rule 1 (candidates only); a bold estimator name marks rule 2's choice for the tail. Every estimator uses year-month FE except Full OLS year FE. Appendix Table A2 reports RMSE, standard-error ratios, and state-clustered coverage by shape.

Figure 3. Simulated error distributions of the cumulative per-day effect: heat (Panel A) and cold (Panel B).

Panel A. Heat (THI $\geq$ 69)



Panel B. Cold (THI $<$ 39)



Note: Error is the estimate minus the true effect, in log points per stress day, cumulated over lags 0–3; 500 replications per shape. Each ridge is scaled to the same peak height, so widths show dispersion; the black tick marks the mean error (bias). Bold: the estimator rule 2 selects for the tail. Groups, top to bottom: OLS; plug-ins; debiased learners.

6 Results

Tables 3 and 4 report the cumulative per-day effects of all seven estimators for milk yield and fat content. Each cell shows the estimate, its two-way standard error in parentheses, and its Driscoll–Kraay standard error in brackets. Bold marks the estimate the decision rules select for each tail: Full OLS for heat and NNET ADML for cold.

6.1 Quantity margin: milk production per cow

All seven estimators give negative cumulative heat and cold responses (Table 3 and Figure 4). An additional heat day lowers milk per cow by 0.289 percent under Full OLS on the cumulative derivative scale, with two-way 95 percent interval [−0.468, −0.110] (Driscoll–Kraay [−0.457, −0.121]). The estimate lies inside the functional-form sensitivity range across the five plausible shapes, −0.219 to −0.363 percent. The debiased learners give larger heat effects: −0.332 percent for UTF-LASSO ADML and −0.327 percent for NNET ADML, 15 and 13 percent above Full OLS. The direction and size of the NNET ADML gap match its overstatement in the simulation. The plug-ins are attenuated (−0.172 and −0.085 percent), also as the simulation predicts, and Linear OLS gives −0.248 percent.

The selected cold estimate is −0.220 percent per day below THI 39, from NNET ADML, with two-way interval [−0.566, 0.126]. The interval includes zero, and the point estimate lies inside the range across shapes, −0.146 to −0.237 percent. Full OLS gives −0.237 percent with interval [−0.451, −0.022], and UTF-LASSO ADML gives −0.255 percent. For cold, the two-way standard errors of Full OLS and the debiased learners are 1.7 to 1.9 times the Driscoll–Kraay ones, against 1.02 to 1.13 times in the simulation. For heat the ratios are 1.00 to 1.14 in the data and 1.07 to 1.14 in the simulation. The simulated response is common to all states. A state-specific component of the actual cold response would enter the two-way variance through its state dimension, while the Driscoll–Kraay variance cannot detect it in a balanced panel. The two-way interval is therefore the cautious choice for cold.

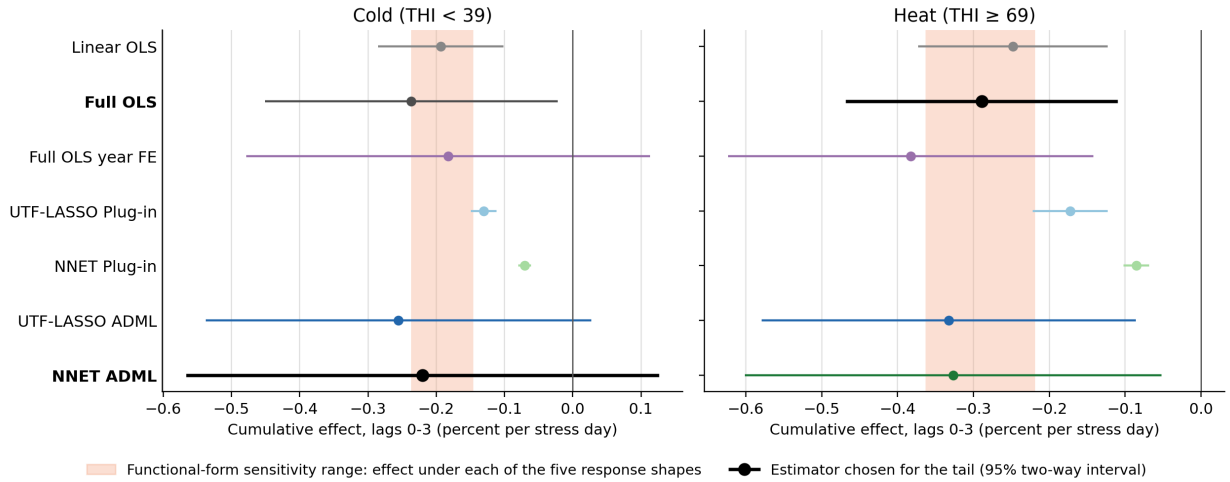
Year FE change the estimates. Full OLS year FE gives a heat effect of −0.383 percent, a third larger than with year-month FE, and a smaller cold effect, −0.183 percent. National monthly shocks correlated with weather therefore affect estimates that control only for year effects; the year-month FE remove them.

Table 3. Cumulative cold- and heat-stress effects on log milk production per cow (`lmpc`), lags 0–3.

Estimator	Cold (THI $<$ 39)	Heat (THI $\geq$ 69)
Linear OLS	−0.00193	−0.00248
	(0.00044) [0.00024]	(0.00060) [0.00051]
Full OLS	−0.00237	**−0.00289**
	(0.00103) [0.00060]	(0.00086) [0.00086]
Full OLS year FE	−0.00183	−0.00383
	(0.00142) [0.00102]	(0.00115) [0.00119]
UTF-LASSO Plug-in	−0.00131	−0.00172
	(0.00009) [0.00006]	(0.00024) [0.00021]
NNET Plug-in	−0.00070	−0.00085
	(0.00005) [0.00003]	(0.00008) [0.00005]
UTF-LASSO ADML	−0.00255	−0.00332
	(0.00135) [0.00075]	(0.00118) [0.00104]
NNET ADML	**−0.00220**	−0.00327
	(0.00166) [0.00089]	(0.00132) [0.00127]

Note: Cumulative average partial derivatives over lags 0–3, in log points per 24-hour-equivalent stress day; 5,712 state-months from 21 states. Two-way (state and sample month) standard errors are in parentheses and Driscoll–Kraay standard errors in brackets. Every estimator includes state × calendar-month FE (with first differencing) and year-month FE, except Full OLS year FE, which uses year FE. Learners use 20 repetitions of five-fold cross-fitting over states. Bold marks the estimator the simulation rules select for the tail. No significance markers are used.

Figure 4. Cumulative heat and cold effects on milk yield across the seven estimators.



Note: Points are cumulative effects over lags 0–3, in percent per stress day (100 times log points); bars are two-way 95 percent intervals. The shaded band is the functional-form sensitivity range: the effect under each of the five response shapes that fit the data (Section 5). Black: the estimator selected for the tail.

6.2 Composition margin: milk fat content

Heat lowers and cold raises milk fat content under all seven estimators (Table 4). The selected heat estimate, from Full OLS, is −0.00452 percentage points per heat day, with two-way interval [−0.00804, −0.00100]. That is about 0.12 percent of the mean fat content of 3.82 percent. The selected cold estimate, from NNET ADML, is 0.01156 percentage points per cold day [0.00466, 0.01846], about 0.30 percent of the mean. NNET ADML gives a much smaller heat effect (−0.00115) because its positive lag-2 and lag-3 contributions are larger than those of Full OLS. A concentration response alone establishes neither a change in total fat output nor a physiological mechanism. The simulation does not cover fat content, so the same estimator per tail is applied without simulation evidence specific to this outcome.

Table 4. Cumulative cold- and heat-stress effects on milk fat content (`fc`), lags 0–3.

Estimator	Cold (THI $<$ 39)	Heat (THI $\geq$ 69)
Linear OLS	0.00685	−0.00476
	(0.00092) [0.00059]	(0.00099) [0.00094]
Full OLS	0.00989	**−0.00452**
	(0.00241) [0.00252]	(0.00169) [0.00225]
Full OLS year FE	0.00772	−0.00444
	(0.00233) [0.00238]	(0.00187) [0.00197]
UTF-LASSO Plug-in	0.00977	−0.00583
	(0.00037) [0.00010]	(0.00026) [0.00012]
NNET Plug-in	0.00606	−0.00493
	(0.00029) [0.00027]	(0.00022) [0.00012]
UTF-LASSO ADML	0.01003	−0.00454
	(0.00275) [0.00287]	(0.00206) [0.00273]
NNET ADML	**0.01156**	−0.00115
	(0.00331) [0.00334]	(0.00303) [0.00349]

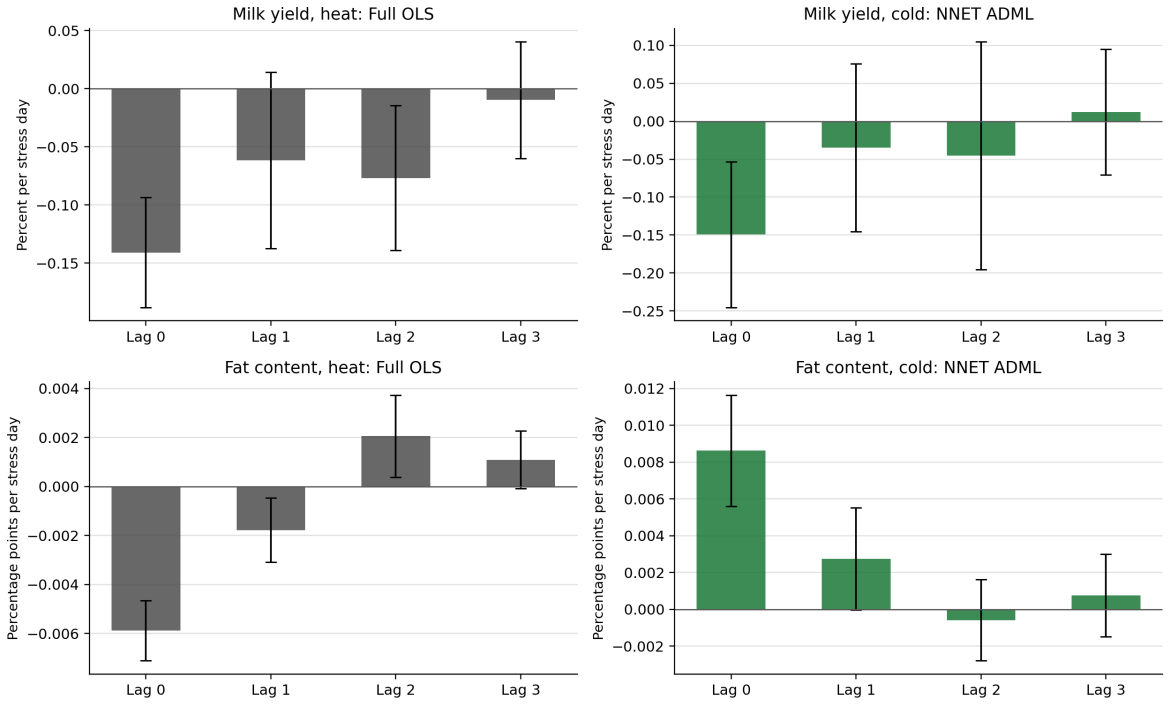
Note: Percentage points of fat content per 24-hour-equivalent stress day, cumulated over lags 0–3; 5,665 state-months from 21 states. Fat content is missing for Colorado in 2003–2004 and for Kansas and Oregon in 2004; differences never span a missing month. Other notes as in Table 3.

6.3 Lag structure

Figure 5 separates the selected estimators' responses into monthly lag contributions. For milk yield, most of the heat response occurs in the exposure month (−0.141 percent, interval [−0.188, −0.094]). Smaller contributions follow at lags 1 and 2 (−0.061 and −0.077 percent; only the latter's interval excludes zero), and little remains by lag 3. The cold response is also concentrated in the exposure month (−0.149 percent [−0.246, −0.053]), and the later lags

have intervals that include zero. For fat content, heat lowers concentration in the exposure month (−0.00589 percentage points) and the next month (−0.00178). Positive contributions at lags 2 and 3 (0.00206 and 0.00109) then partly reverse the decline. Cold raises fat content mainly in the exposure month (0.00862). Separate lag intervals describe the distributed response; they are not a joint test of persistence, which Section 4.4 reports for lags 4–6.

Figure 5. Lag-specific effects of the selected estimators.



Note: Heat panels use Full OLS and cold panels NNET ADML, the estimators selected for each tail. Yield effects are in percent per stress day (100 times log points) and fat effects in percentage points; bars are two-way 95 percent intervals. Year-month FE throughout.

6.4 Intensity-duration measures

Appendix Table A1 repeats the estimates with the shifted-log load measures, which weight each hour by its distance beyond the threshold. Five of the seven estimators give negative yield effects for both tails. The selected estimators give −0.00914 log points per unit of heat load (Full OLS, interval [−0.01348, −0.00480]) and −0.00705 per unit of cold load (NNET ADML, [−0.01132, −0.00277]). Where load is large relative to 0.001, a 1 percent increase in heat load corresponds to an approximately 0.009 percent decline in milk per cow on the cumulative derivative scale. Linear OLS has the opposite sign for both tails (0.00041 for heat and 0.00100 for cold), which shows that a linear term in log load misrepresents

the response. NNET Plug-in recovers almost none of the load response (estimates near zero and positive), so the debiasing correction carries the NNET ADML load estimates. For fat content, heat load lowers and cold load raises concentration under the selected estimators (-0.01274 and 0.02103 percentage points). No simulation evidence covers the load measures.

7 Temporal variation and adaptation diagnostics

7.1 Annual averages of the scores

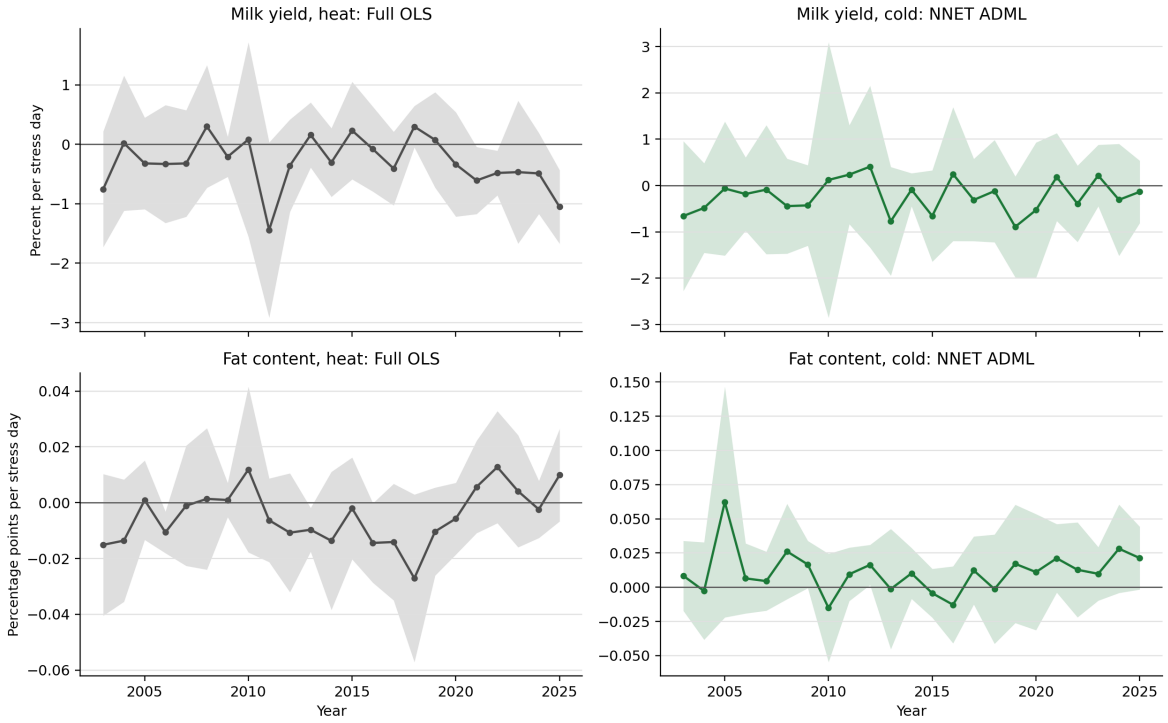
For outcome k, stress tail r, and year v, we average the cumulative scores of the selected estimator over the year's state-months. For NNET ADML these are the held-out scores; for Full OLS, the influence-function scores of the full-sample fit:

$$\hat{\delta}_{k,r,v} = \frac{1}{N_{k,v}} \sum_{(i,t):\,\mathrm{year}(t)=v} \sum_{\ell=0}^{3} \widehat{\psi}_{kr\ell,it}. \tag{6}$$

The response is not refitted by year. Annual averages can move with the year's exposure distribution and residual correction even when the pooled response is constant, so they describe temporal variation rather than identify year-specific effects. Intervals use the two-way variance within the year's state-months, and NNET ADML averages are aggregated over the 20 fold assignments as in Section 4.3.

Figure 6 shows the annual series. Comparing 2016–2025 with 2006–2015, which gives complete calendars for both outcomes, the yield heat average changes by -0.136 percent per day (two-way SE 0.175) and the yield cold average by -0.029 percent (0.231). The fat-content averages change by -0.00017 percentage points for heat (0.00439) and 0.00503 for cold (0.00631). None of the four intervals excludes zero, so the decade averages do not differ detectably. Because they summarize the pooled scores rather than estimate decade-specific responses, this comparison is not a test of a change in the response itself; the break search in Section 7.2 tests that directly.

Figure 6. Annual averages of cumulative scores for milk yield and fat content (2003–2025).



Note: Each point averages the selected estimator's cumulative scores (lags 0–3, year-month FE) over the year's state-months: held-out scores for NNET ADML (cold) and full-sample influence-function scores for Full OLS (heat). Shading shows pointwise two-way 95 percent intervals; it is not a trend or break test. 2003 begins in May after lag construction, and fat content is missing for three states in 2003–2004.

7.2 Break search

The annual averages describe the scores; a break search asks whether the response itself changed. For each candidate year b from 2008 to 2020, we refit Full OLS with every weather coefficient allowed to differ after year b. The change in the cumulative per-day effect evaluates the responses before and after b at the full sample's exposure, so a shift in the exposure distribution does not enter. A sup test takes the largest absolute t-statistic over the 13 candidate years. Its critical values are simulated from the estimated correlation of the 13 changes, with $t(20)$ tails for the two-way variance and normal tails for Driscoll–Kraay. Appendix Table A6 reports every change.

Neither yield response changes detectably. The largest heat change comes after 2019: the effect is −0.390 percent per day afterward against −0.211 before, a change of −0.179 (two-way SE 0.070), but the sup-test p-value is 0.087 (0.132 with Driscoll–Kraay). For cold it is 0.160. Fat content behaves differently. The heat effect on fat content weakens after the

mid-2010s: with the break after 2016 it is −0.00692 percentage points per day before and −0.00008 after, and the sup test rejects a stable response ($p < 0.001$ with either variance). The cold effect on fat content grows at the end of the sample (largest change after 2020, sup-test $p = 0.014$).

The search has limits. It uses Full OLS for both tails, because refitting NNET ADML for every candidate year is costly, and the simulation does not cover this specification. The state × calendar-month FE are fixed over time, so a change in the seasonal pattern of fat content that coincides with hot or cold months would also appear as a change in the response. With 21 states, the search also has little power against moderate changes in the yield response.

7.3 Climate-normal heterogeneity

A separate diagnostic asks whether yield responses to unusual weather differ with a state's climate history. It regresses log milk yield in levels on heat and cold shocks, the trailing state × calendar-month normals of each exposure (centered), precipitation shocks and normals, and shock-by-normal interactions, with lags 0–3. Normals average the same calendar month over the preceding ten years, so they use only past weather. The specification includes state × calendar-month FE and year-month FE. With ten-year normals the sample has 3,213 state-months from 21 states. This OLS diagnostic has a different specification and sample from the main estimates.

A heat-shock day lowers yield by 0.377 percent at the average normal (two-way SE 0.137). The cumulative heat interaction is 0.000305 (two-way SE 0.000104; interval [0.000088, 0.000522]): each additional day of normal monthly heat exposure makes the response to a heat-shock day 0.031 percentage points less negative. The cold interaction is close to zero (0.000033, SE 0.000095). The heat interaction is not robust to the choice of windows, however. Across four normal windows (5 to 15 years) and four lag windows (0–1 to 0–4), its two-way interval lies above zero in only 4 of the 16 combinations (Appendix Table A5). The diagnostic is therefore suggestive evidence that climate history is associated with smaller heat damage. It does not identify the effect of adaptation investments or an adaptation rate.

8 Economic magnitude

8.1 Method and scope

The production share (3) is the production-weighted mean change in log yield from removing a stress tail, at the observed exposure of every state-month. We estimate it with the selected estimators: Full OLS for heat and NNET ADML for cold. Removing a tail replaces its stress days with thermoneutral exposure in the current and three preceding months, holding the other tail and precipitation fixed. The joint share is the sum of the heat and cold shares, with standard errors from the sum of their scores within each fold assignment. The sum equals the effect of removing both tails together only if the tails act additively on yield. Full OLS imposes this, and its own joint share, 2.91 percent, is close to the sum of 2.89 percent. NNET ADML lets the tails interact, yet within each fold assignment its joint share differs from the sum of its heat and cold shares by at most 0.0032 percentage points; its joint share is larger, 3.24 percent, because its heat share is larger (Appendix Table A7). We report shares as 100 times log points ("percent"); at the magnitudes in Table 5 they differ from proportional changes by less than 2.5 percent of their value.

Revenue equivalents are an accounting that applies one share to every state and year. They convert the share into lost production, $Q_{iy}\{\exp(\hat{\theta})-1\}$, using NASS annual production Q_{iy} by state and year, and value it at the state's annual all-milk price P_{iy} in dollars per hundredweight. The share is estimated over May 2003 through December 2025 and applied to each year of 2003–2024; evaluating the Full OLS response over the 2003–2024 months alone changes its heat and cold shares by less than 0.01 percentage points (Appendix F). The estimation sample covers the 21 states, which produced 91.3 percent of contiguous-U.S. milk in 2003–2024. Extending the share to the 48 contiguous states assumes that the remaining production faces a similar response and exposure. We report both totals for 2003–2024, with the average per year, total divided by 22, in parentheses. The calculation holds prices, herd size, other inputs, and producer behavior fixed, so it measures gross revenue equivalents rather than profits, welfare, or equilibrium damages.

8.2 Production shares and revenue equivalents

The heat share is 1.68 percent (log points; two-way interval 0.53 to 2.82), within the range across shapes of 1.28 to 2.12 percent (Table 5). For the 48 contiguous states this is

$13.78 billion over 2003–2024 ($0.63 billion per year), with interval $4.34 to $23.32 billion. The cold share is 1.21 percent (interval −0.22 to 2.65; shape range 0.98 to 1.37), or $9.93 billion ($0.45 billion per year). The cold interval includes zero, as does its per-day effect. The joint share is 2.89 percent (interval 1.21 to 4.57; shape range 2.26 to 3.50), or $23.88 billion ($1.09 billion per year), with interval $9.91 to $38.09 billion. For the 21 estimation states alone the totals are $12.53 billion for heat ($0.57 billion per year) and $21.72 billion for both tails ($0.99 billion per year).

Valuing every year at each state's 2024 price, which holds the price vector fixed rather than adjusting for inflation, the joint total for the 48 states is $29.48 billion ($1.34 billion per year). In 2024 alone, heat stress cost $0.86 billion and the two tails together $1.49 billion. Appendix Table A7 reports each estimator's own shares. NNET ADML's heat share, 2.09 percent, is 24 percent above Full OLS, consistent with its simulated heat overstatement. The load measures give a heat share of 1.32 percent and a joint share of 2.07 percent, and Full OLS year FE gives 1.70 percent for heat.

Evaluated at each state's own exposure, the Full OLS heat share ranges from 0.72 percent in Oregon to 4.54 percent in Florida (Appendix Table A8 and Figure A1). It exceeds 2 percent in Florida, Arizona, Texas, New Mexico, Kansas, and California. California has the largest heat revenue equivalent, $3.23 billion over 2003–2024 ($147 million per year), followed by Wisconsin and Texas ($1.33 and $1.32 billion). The state values sum to $12.50 billion, close to the $12.53 billion from applying one share to every state. They rest on the common response, so they differ only through exposure, not through state differences in sensitivity such as cooling or management. In Florida, Arizona, Texas, and California, no four-month history in the sample is free of heat days, so their shares extrapolate the common response to zero exposure. Their intervals hold each state's exposure fixed. The cold shares, also from Full OLS, are reported without intervals because the simulation does not support Full OLS intervals for cold.

Table 5. Production shares and revenue equivalents of heat and cold stress, 2003–2024.

	Production share (%)			Revenue equivalent (billion $)		
Stress removed	Estimate	95% interval	Shape range	21 states	48 states	95% interval (48)
Heat (Full OLS)	1.68	[0.53, 2.82]	1.28–2.12	12.53 (0.57)	13.78 (0.63)	[4.34, 23.32]
Cold (NNET ADML)	1.21	[−0.22, 2.65]	0.98–1.37	9.04 (0.41)	9.93 (0.45)	[−1.82, 21.86]
Heat and cold (sum)	2.89	[1.21, 4.57]	2.26–3.50	21.72 (0.99)	23.88 (1.09)	[9.91, 38.09]

Note: The production share is the production-weighted mean change in log milk per cow when the tail is removed at lags 0–3 (year-month FE; heat from Full OLS, cold from NNET ADML, the joint share their sum, which assumes the tails act additively). Intervals are two-way, with $t(20)$ critical values. The shape range is the functional-form sensitivity range: the share under each of the five response shapes that fit the data (Section 5). Revenue equivalents apply the share uniformly to every state and year: $Q\{\exp(\theta) - 1\}$ valued at state annual all-milk prices, in nominal dollars, summed over 2003–2024; the average per year (total divided by 22) is in parentheses. The 48-state column applies the estimated share to all contiguous states.

The fat-content results add a composition margin that this all-milk valuation leaves out. Heat lowers and cold raises fat concentration, but concentration changes alone do not identify changes in total fat output or milk value. Valuing them would require a component-pricing framework and a compatible joint quantity–composition counterfactual.

9 Discussion and conclusion

9.1 Findings and implications

Our estimates place both heat and cold among the production risks facing U.S. dairy systems. The negative responses in both tails agree qualitatively with the two-sided pattern documented by Gisbert-Queral et al. (2021) and Choi et al. (2024), whose temperature-bin contrasts and response curves differ from our cumulative derivatives per equivalent stress day; our heat boundary of THI 69 also includes moderate exposure, while the headline extreme-heat contrast of Gisbert-Queral et al. concerns THI above 79. The heat estimate rules out a zero effect; the cold estimate does not, and its size depends on the shape of the response.

Heat stress reduced production by about 1.7 percent over the sample period (1.3 to 2.1 percent across plausible shapes). Mauger et al. (2015) estimate a historical national loss of 1.9 percent, but against an idealized no-heat-stress baseline with different climate years, cow populations, and response model, so the similar magnitudes do not validate each other. Our joint share of 2.9 percent adds cold stress to heat-focused assessments. Across states, the

heat share is largest in Florida and the Southwest, where heat is most common, but those shares also rely most on extrapolation.

Heat lowers and cold raises fat concentration, consistent with the importance of composition documented at the cow and farm levels by Prieto et al. (2026) and Hutchins et al. (2025). Concentration differs from total component output, however, because fat production depends jointly on volume and concentration, so positive cold concentration responses cannot be read as an offset to quantity losses without a component calculation.

Temporal evidence also depends on what each study measures. Choi et al. (2024) report lower yield sensitivity in recent decades, whereas Prieto et al. (2026) find broadly similar heat-response curves for 2007–2011 and 2012–2016. Our break search finds no detectable change in the yield responses, but the heat effect on fat content weakens after the mid-2010s. The climate-normal diagnostics point to smaller heat damage where heat is historically common, but that pattern is sensitive to specification, and our state aggregates cannot identify the management heterogeneity suggested by the small-herd vulnerability that Hutchins et al. document.

The methodological results bear on how flexible estimators should be used in climate-impact panels. Automatic debiasing did not uniformly improve on a transparent cubic regression. For heat, Full OLS had the smallest worst-case bias, and the debiased neural network overstated the effect under every shape, because its first stage recovered only about a third of the effect and the correction inherited the error of an approximate representer. The debiased LASSO of Klosin and Vilgalys (2025) performed about as well as Full OLS for heat once the time FE were left unpenalized, which keeps them as unrestricted additive effects, as in their model, so no part of a national shock is shrunk back into the regression; its plug-in, by contrast, was badly biased. For cold, no estimator was robust to the response shape. Rules established before the simulation discipline these comparisons, but simulated rankings still depend on whether biases and variances transfer to the application (Advani et al., 2019), which is why we anchor the simulation to the observed weather, sample, and residual dependence and report the range across shapes.

9.2 Limitations

State aggregation hides herd, breed, housing, and management heterogeneity. Causal interpretation requires the exogeneity condition of Section 4.1. Year-month FE remove national shocks, but regional shocks that coincide with regional weather can remain; we do not estimate region × month FE. The exposure weights come from the nearest agricultural census, which for some years is a later one, so they can partly reflect where herds moved afterward. The per-day effects and production shares vary by 40 to 66 percent across response shapes that fit the data equally well, which is why we report ranges, and the cold estimate is fragile. The production shares remove each tail completely, but a four-month history without heat days occurs in only 50 of the 252 state × calendar-month cells (36 for cold), so in the other cells the shares rely on the common response to extrapolate to zero exposure. They also cover lags 0–3 only, and longer windows give larger but less precise cumulative effects (Section 4.4). The simulation evaluates the per-day yield effect with one fold assignment per replication; the medians over 20 assignments, the production shares, and the fat-content and load results rest on the same estimators without simulation evidence of their own. Two-way and Driscoll–Kraay inference allow correlation across states, and the cross-fitting assumes that this correlation is weak once the year-month FE are removed (Section 4.3); we do not estimate spatial models of the kind used by Fan and Ma (2024). The valuation assumes that the estimated share applies in every year and to the 8.7 percent of production outside the estimation states, and it holds prices, herd size, and behavior fixed. It measures gross revenue equivalents, not welfare, and omits fat-component revenue. The state shares reflect exposure under the common response, not state differences in sensitivity, and the hottest states extrapolate that response to zero exposure. The break search uses Full OLS for both tails and cannot separate a change in the stress response from a change in the seasonal pattern of fat content. None of the temporal diagnostics identifies adaptation.

9.3 Concluding remarks

The study estimates the effects of both stress tails on two production margins in a design that absorbs national monthly shocks. It chooses its estimators by rules established before a synthetic-data simulation. Heat stress has a clear negative effect on milk yield, about 0.29 percent per additional heat day and 1.7 percent of production over 2003–2024.

Cold stress is estimated with much less certainty. The size of both effects depends on a response shape the data cannot pin down, so ranges across plausible shapes are a more honest summary than any single functional form. Further work should connect herd-level responses to observed management, explain why the heat effect on fat content weakened after the mid-2010s, combine volume and component output in a pricing framework, and develop designs that separate regional shocks from regional weather.

Appendix

Appendix A. Estimation details

Scaling and outcome learners

Within each outer training fold, the level basis is standardized by its training-sample standard deviations; the derivative directions use the same scales. The transformation of Section 4.2 is applied to every column.

UTF-LASSO partials the FE out of the outcome and the weather columns by least squares in the outer training sample. It then fits a LASSO to the residualized weather columns, with the penalty chosen by five-fold state-grouped cross-validation over 50 candidates (at most 20,000 iterations, tolerance 10^{-5}), and recovers the FE coefficients by least squares. Its plug-in derivative is $d'_{r\ell,it}\hat{\beta}$, where $d_{r\ell,it}$ is the derivative of the scaled basis with respect to $D^r_{i,t-\ell}$.

NNET maps the scaled level weather history x_{it} (both exposures and precipitation at lags 0–3, plus the sine and cosine of the calendar month) to a level response $f(x_{it})$ through two hidden layers of 64 and 32 units with CELU activation. The FE enter through a separate linear layer without bias, and an unpenalized constant in the differenced prediction absorbs a common drift. The fitted transformed prediction is the transformation of $f(x_{it}) - f(x_{i,t-1})$ plus the FE term, and training minimizes its mean squared error with full-batch Adam (learning rate 0.01). Weight decay of 10^{-4} applies to f only. Training stops early on held-out training states (about one fifth of them), with patience 100 and at most 3,000 epochs. The plug-in derivative $\partial f(x_{it})/\partial D^r_{i,t-\ell}$ is computed by automatic differentiation.

Riesz representer

For training sample $\mathcal{T}$ of size $N_{\mathcal{T}}$, with transformed basis Z and derivative direction $d_{r\ell}$, let

$$\widehat{Q} = N_{\mathcal{T}}^{-1} \sum_{\mathcal{T}} Z_{it} Z_{it}', \qquad \widehat{M}_{r\ell} = N_{\mathcal{T}}^{-1} \sum_{\mathcal{T}} d_{r\ell,it}. \tag{A1}$$

The representer coefficients solve

$$\hat{\rho}_{r\ell} = \arg\min_{\rho} \left\{ \rho' \widehat{Q} \rho - 2\widehat{M}_{r\ell}' \rho + c_{r\ell} \lambda \|\rho\|_1 \right\}, \qquad \lambda = \text{lnln}(N_{\mathcal{T}}) \sqrt{\frac{\ln p}{N_{\mathcal{T}}}}. \tag{A2}$$

The multiplier $c_{r\ell}$ is chosen from 13 values, log-spaced from 10 to 0.001, to minimize the held-out-state Riesz loss $\rho' \widehat{Q}_v \rho - 2\widehat{M}_{r\ell,v}' \rho$ over five state-grouped inner folds. The FE columns are partialled out of Z before (A2). Every derivative direction is zero on the FE columns, so this is equivalent to leaving their coefficients unpenalized. For NNET ADML, Z also includes natural cubic spline terms in each exposure and lag. Their five knots are the exposure floor, the three quartiles of exposure above it, and the maximum, which adds three nonlinear columns per exposure and lag. The production share (3) uses the direction $w_{it}\{b(W_{it}^r) - b(W_{it})\}$ in place of $d_{r\ell,it}$.

Variance estimators

The two-way variance of a score mean adds the state- and month-clustered variances, each with its $G/(G-1)$ factor, and subtracts the heteroskedasticity-robust variance (Cameron et al., 2011); if the result is not positive, the larger one-way variance is used. The Driscoll–Kraay variance applies a Bartlett kernel with a 12-month bandwidth to the scores summed within each calendar month (Driscoll & Kraay, 1998). The state-clustered comparison is the state term alone. Lag scores are summed before these variances are computed, which retains their covariance.

Appendix B. Additional estimates

Table A1. Cumulative effects of the shifted-log intensity-duration (load) measures, lags 0–3.

Estimator	Cold load	Heat load
Panel A. Milk yield		
Linear OLS	0.00100	0.00041
	(0.00067) [0.00058]	(0.00045) [0.00044]
Full OLS	−0.00591	**−0.00914**
	(0.00180) [0.00098]	(0.00208) [0.00181]
Full OLS year FE	−0.00512	−0.00667
	(0.00223) [0.00159]	(0.00210) [0.00196]
UTF-LASSO Plug-in	−0.00441	−0.00561
	(0.00046) [0.00018]	(0.00071) [0.00036]
NNET Plug-in	0.00010	0.00006
	(0.00003) [0.00003]	(0.00002) [0.00001]
UTF-LASSO ADML	−0.00594	−0.00995
	(0.00201) [0.00120]	(0.00258) [0.00222]
NNET ADML	**−0.00705**	−0.01262
	(0.00205) [0.00153]	(0.00365) [0.00291]
Panel B. Fat content		
Linear OLS	0.00482	−0.00246
	(0.00123) [0.00115]	(0.00133) [0.00132]
Full OLS	0.01814	**−0.01274**
	(0.00284) [0.00227]	(0.00371) [0.00337]
Full OLS year FE	0.01382	−0.02348
	(0.00319) [0.00254]	(0.00350) [0.00291]
UTF-LASSO Plug-in	0.01654	−0.01110
	(0.00103) [0.00026]	(0.00073) [0.00023]
NNET Plug-in	0.00437	−0.00308
	(0.00041) [0.00026]	(0.00046) [0.00036]
UTF-LASSO ADML	0.01840	−0.01257
	(0.00334) [0.00271]	(0.00436) [0.00388]
NNET ADML	**0.02103**	−0.01503
	(0.00414) [0.00335]	(0.00634) [0.00463]

Note: Effects per unit of ln(Load + 0.001), in log points (yield) and percentage points (fat). Two-way standard errors in parentheses and Driscoll–Kraay in brackets. Bold marks the estimator selected for the tail on the stress-day measures. The simulation does not cover the load measures.

Appendix C. Simulation details

Table A2. Simulation results by estimator and response shape.

				SE / SD			Coverage		
Estimator	Shape	Bias (%)	RMSE	Two-way	DK	State	Two-way	DK	State
Panel A. Heat									
Linear OLS	Cubic	+15.8	0.712	0.84	0.76	0.68	0.814	0.728	0.640
	Convex	−11.1	0.597	0.84	0.76	0.69	0.904	0.802	0.758
	Natural spline	+3.6	0.553	0.84	0.76	0.68	0.916	0.854	0.814
Full OLS	Cubic	+2.6	0.955	0.98	0.91	0.83	0.956	0.928	0.886
	Convex	+3.1	0.955	0.98	0.91	0.83	0.960	0.928	0.884
	Natural spline	−9.9	0.984	0.98	0.91	0.83	0.946	0.906	0.864
Full OLS year FE	Cubic	+1.6	1.107	1.03	0.96	0.72	0.968	0.946	0.810
	Convex	+1.1	1.106	1.03	0.96	0.72	0.968	0.946	0.822
	Natural spline	−12.2	1.149	1.03	0.96	0.72	0.956	0.926	0.798
UTF-LASSO Plug-in	Cubic	+17.6	0.858	0.15	0.02	0.15	0.174	0.022	0.156
	Convex	+16.4	0.682	0.24	0.03	0.23	0.286	0.026	0.258
	Natural spline	+7.0	0.707	0.15	0.02	0.15	0.224	0.026	0.216
NNET Plug-in	Cubic	+70.1	2.039	0.16	0.00	0.16	0.000	0.000	0.000
	Convex	+60.5	1.346	0.16	0.00	0.16	0.000	0.000	0.000
	Natural spline	+65.9	1.680	0.16	0.00	0.16	0.000	0.000	0.000
UTF-LASSO ADML	Cubic	+2.4	1.160	1.01	0.89	0.88	0.970	0.924	0.914
	Convex	+2.1	1.161	1.02	0.89	0.88	0.974	0.928	0.920
	Natural spline	−10.4	1.189	1.01	0.89	0.88	0.966	0.912	0.890
NNET ADML	Cubic	−9.2	1.452	1.04	0.96	0.87	0.974	0.942	0.902
	Convex	−8.7	1.444	1.04	0.95	0.87	0.974	0.938	0.900
	Natural spline	−14.3	1.471	1.04	0.96	0.87	0.968	0.936	0.908
Panel B. Cold									
Linear OLS	Cubic	+17.4	0.546	0.97	0.94	0.72	0.790	0.722	0.598
	Convex	−36.0	0.635	0.96	0.94	0.72	0.684	0.638	0.486
	Natural spline	−25.3	0.533	0.97	0.94	0.72	0.806	0.750	0.598
Full OLS	Cubic	−1.0	0.880	0.93	0.91	0.76	0.944	0.926	0.856
	Convex	−6.8	0.885	0.93	0.91	0.76	0.936	0.922	0.842
	Natural spline	−54.9	1.228	0.93	0.91	0.76	0.816	0.780	0.690
Full OLS year FE	Cubic	−1.8	1.113	0.96	0.93	0.62	0.940	0.892	0.792
	Convex	−8.0	1.118	0.96	0.93	0.62	0.934	0.892	0.788
	Natural spline	−51.8	1.375	0.96	0.93	0.62	0.900	0.862	0.666
UTF-LASSO Plug-in	Cubic	+28.6	0.850	0.15	0.02	0.15	0.118	0.012	0.110
	Convex	+9.2	0.438	0.19	0.03	0.18	0.280	0.050	0.240
	Natural spline	−8.2	0.529	0.16	0.02	0.15	0.258	0.034	0.240
NNET Plug-in	Cubic	+68.2	1.624	0.15	0.00	0.15	0.000	0.000	0.000
	Convex	+47.7	0.717	0.15	0.00	0.15	0.000	0.000	0.000
	Natural spline	+51.9	0.829	0.15	0.00	0.15	0.000	0.000	0.000
UTF-LASSO ADML	Cubic	−2.1	1.013	0.96	0.91	0.78	0.942	0.918	0.854
	Convex	−4.8	1.021	0.95	0.90	0.77	0.938	0.922	0.868
	Natural spline	−57.5	1.354	0.95	0.91	0.78	0.864	0.818	0.730
NNET ADML	Cubic	−5.2	1.214	1.07	0.95	0.90	0.964	0.922	0.898
	Convex	−2.2	1.206	1.04	0.95	0.86	0.962	0.922	0.884
	Natural spline	−46.4	1.408	1.07	0.95	0.89	0.930	0.882	0.828

Note: 500 replications per shape. Bias is relative to the absolute true effect; RMSE is in 10^{-3} log points per stress day. SE / SD is the mean reported standard error over the standard deviation of the estimates. The noise transformations come from a finite set, and replications after the 100th use transformations not drawn before, so binomial Monte Carlo standard errors of the coverage rates are conservative for this design.

Table A3. First-tier test of the response shapes and their exact true effects.

		Percentile			Per-day effect (%)		Production share (%)		
Shape	Fit R^2	Heat FP	Cold FP	Passes	Heat	Cold	Heat	Cold	Joint
Linear	0.3998	64.4	75.0	Yes	−0.263	−0.184	1.55	1.20	2.75
Cubic	0.4098	51.2	51.8	Yes	−0.289	−0.237	1.68	1.23	2.91
Convex	0.4093	79.4	85.4	Yes	−0.219	−0.146	1.28	0.98	2.26
Concave	0.4059	27.8	48.2	Yes	−0.363	−0.230	2.12	1.37	3.50
Natural spline	0.4117	53.8	50.4	Yes	−0.252	−0.156	1.61	1.11	2.73

Note: The fingerprints are the differences between the Linear and Full OLS cumulative effects for heat and cold. Percentiles locate their observed values in 500 simulated replications per shape; a shape passes if both lie between the 2.5th and 97.5th percentiles. Per-day effects are 100 times log points per stress day; production shares are 100 times log points.

Appendix D. Lag window

Table A4. Lag-window checks: stopping lags of the sequential lag procedure and cumulative effects of the selected estimators by window.

	Cold			Heat		
Estimator	Two-way	DK	State	Two-way	DK	State
Panel A. Milk yield						
Linear OLS	1	2	2	0	0	0
Full OLS	1	1	1	0	0	0
UTF-LASSO ADML	1	1	1	0	0	0
NNET ADML	0	0	0	0	0	0
Panel B. Fat content						
Linear OLS	1	1	1	1	1	3
Full OLS	1	1	1	1	1	3
UTF-LASSO ADML	1	1	1	1	1	1
NNET ADML	1	1	1	3	4	5

	Milk yield		Fat content	
Window	Cold	Heat	Cold	Heat
Panel C. Cumulative effect by window				
Lags 0–3	−0.00201	−0.00283	0.01148	−0.00401
	(0.00152)	(0.00090)	(0.00347)	(0.00171)
Lags 0–4	−0.00224	−0.00352	0.01162	−0.00455
	(0.00164)	(0.00147)	(0.00343)	(0.00237)
Lags 0–5	−0.00361	−0.00462	0.01562	0.00002
	(0.00268)	(0.00147)	(0.00476)	(0.00308)
Lags 0–6	−0.00515	−0.00431	0.01725	−0.00169
	(0.00319)	(0.00171)	(0.00552)	(0.00426)
Lags 4–6 within 0–6	−0.00222	−0.00040	0.00285	0.00112
	(0.00153)	(0.00092)	(0.00275)	(0.00213)
Two-way p-value	0.16	0.67	0.31	0.61

Note: Windows 0–0 to 0–12 are estimated on fixed samples (yield 5,523 and fat 5,487 state-months over 263 months) with year-month FE and a single fold assignment. Panels A and B: each entry is the last lag before the first newly added lag whose 95 percent interval includes zero. The rule uses the two-way intervals and keeps lags 0–3 when every stopping lag is 3 or less. Panel C: cumulative effects of the selected estimators (Full OLS for heat, NNET ADML for cold) over each window, in log points per stress day for yield and percentage points for fat content, with two-way standard errors in parentheses. Because the samples are fixed and NNET ADML uses one fold assignment, the 0–3 values differ slightly from Tables 3 and 4. The last rows give the sum of the lag 4–6 effects within the 0–6 window and its two-way p-value, with $t(20)$ critical values. Of the estimators not selected for a tail, only Full OLS for fat-content cold has a lag 4–6 sum significant at 5 percent ($p = 0.048$).

Appendix E. Temporal and adaptation diagnostics

Table A5. Climate-normal heat interaction across normal and lag windows.

Normal window (years)	Lags	Year-month FE	Year FE	N
5	0–1	−1.05 (0.90)	−1.23 (0.79)	4,515
5	0–2	−0.83 (1.15)	−0.94 (1.02)	4,494
5	0–3	−0.90 (1.39)	−1.08 (1.23)	4,473
5	0–4	−1.17 (1.60)	−0.98 (1.46)	4,452
7	0–1	−0.42 (0.93)	−0.75 (0.81)	4,011
7	0–2	0.11 (1.17)	−0.34 (1.10)	3,990
7	0–3	0.41 (1.43)	−0.07 (1.33)	3,969
7	0–4	0.35 (1.70)	0.41 (1.61)	3,948
10	0–1	1.05 (0.76)	0.95 (0.67)	3,255
10	0–2	**2.46 (0.88)**	**2.69 (0.89)**	3,234
10	0–3*	**3.05 (1.04)**	**3.42 (1.10)**	3,213
10	0–4	**3.20 (1.22)**	**3.89 (1.36)**	3,192
15	0–1	1.04 (0.89)	0.30 (0.70)	1,995
15	0–2	**2.45 (1.06)**	1.72 (1.03)	1,974
15	0–3	2.63 (1.36)	1.91 (1.39)	1,953
15	0–4	2.67 (1.56)	2.41 (1.57)	1,932

Note: Cumulative heat-shock × normal interaction $\times 10^4$, with two-way standard errors ($\times 10^4$) in parentheses. Both specifications include state × calendar-month FE. Bold: the two-way 95 percent interval lies above zero. *Baseline specification.

Table A6. Break search: change in the cumulative per-day effect after each candidate year.

	Milk yield (percent)		Fat content (percentage points)	
Break after	Heat	Cold	Heat	Cold
2008	0.018	0.009	0.00434	−0.00092
	(0.132)	(0.150)	(0.00206)	(0.00428)
2009	−0.064	0.048	0.00538	−0.00160
	(0.132)	(0.115)	(0.00160)	(0.00365)
2010	−0.052	−0.018	0.00486	−0.00115
	(0.132)	(0.112)	(0.00195)	(0.00400)
2011	0.045	−0.038	0.00275	0.00227
	(0.128)	(0.098)	(0.00156)	(0.00338)
2012	0.115	−0.119	0.00529	0.00189
	(0.118)	(0.087)	(0.00230)	(0.00337)
2013	0.080	−0.152	0.00383	0.00132
	(0.136)	(0.087)	(0.00161)	(0.00318)
2014	−0.057	−0.208	0.00402	0.00103
	(0.136)	(0.093)	(0.00121)	(0.00348)
2015	−0.271	0.034	0.00412	0.00163
	(0.109)	(0.137)	(0.00173)	(0.00296)
2016	−0.092	−0.031	0.00683	0.00268
	(0.108)	(0.107)	(0.00109)	(0.00272)
2017	−0.133	−0.125	0.00442	0.00502
	(0.082)	(0.098)	(0.00078)	(0.00286)
2018	−0.054	−0.071	0.00653	0.00459
	(0.111)	(0.087)	(0.00225)	(0.00197)
2019	−0.179	−0.040	0.00655	0.00689
	(0.070)	(0.091)	(0.00138)	(0.00216)
2020	−0.175	−0.078	0.00966	0.00868
	(0.083)	(0.085)	(0.00171)	(0.00251)
Largest $\lvert t\rvert$ (after year)	2.58 (2019)	2.25 (2014)	6.25 (2016)	3.45 (2020)
Sup-test p-value, two-way	0.087	0.160	<0.001	0.014
Sup-test p-value, DK	0.132	0.139	<0.001	0.014

Note: Each entry refits Full OLS (year-month FE, lags 0–3) with every weather coefficient allowed to change after the year in the first column and reports the change in the cumulative per-day effect, after minus before, with both responses evaluated at the full sample's exposure; two-way standard errors in parentheses. Before the break, heat lowers yield and fat content and cold raises fat content, so a negative yield change means a larger loss, and a positive fat change means a smaller heat reduction or a larger cold increase. The sup test takes the largest absolute t-statistic over the 13 years; its critical values are simulated from the estimated correlation of the 13 changes, with $t(20)$ tails for the two-way variance and normal tails for Driscoll–Kraay (DK). The largest $\lvert t\rvert$ row uses the two-way variance.

Appendix F. Valuation details

The production share averages log changes across state-months. The exact proportional gain in production, $\sum w\{e^d - 1\}/\sum w$, is at least $e^\theta - 1$ because the exponential is convex. For the Full OLS stress-day shares the three measures are close: 1.68 percent (log points),

1.69, and 1.71 percent for heat; 1.23, 1.24, and 1.25 percent for cold; and 2.91, 2.95, and 2.96 percent for both tails. The revenue equivalents, which use $e^{\theta} - 1$, therefore understate the exact gain by about 1 percent of its value or less. Evaluating the same Full OLS response over the 2003–2024 months alone gives shares of 1.67 percent for heat and 1.23 percent for cold.

Table A7. Production shares and revenue equivalents by estimator, measure, and fixed-effect specification.

				Revenue (billion $)	
Estimator	Removed	Share (%)	Plug-in (%)	Nominal	2024 prices
Panel A. Stress days, year-month FE					
Full OLS	Heat	1.68 (0.55) [0.34]	–	13.78 (0.63)	17.01 (0.77)
Full OLS	Cold	1.23 (0.55) [0.22]	–	10.09 (0.46)	12.46 (0.57)
Full OLS	Joint	2.91 (0.78) [0.40]	–	24.04 (1.09)	29.68 (1.35)
NNET ADML	Heat	2.09 (0.73) [0.45]	0.48	17.17 (0.78)	21.20 (0.96)
NNET ADML	Cold	1.21 (0.69) [0.31]	0.45	9.93 (0.45)	12.26 (0.56)
NNET ADML	Joint	3.24 (0.75) [0.51]	0.93	26.81 (1.22)	33.10 (1.50)
Full OLS + NNET ADML	Joint	2.89 (0.81) [0.42]	–	23.88 (1.09)	29.48 (1.34)
Panel B. Stress load, year-month FE					
Full OLS	Heat	1.32 (0.58) [0.43]	–	10.80 (0.49)	13.34 (0.61)
Full OLS	Cold	0.73 (0.65) [0.55]	–	5.94 (0.27)	7.34 (0.33)
Full OLS	Joint	2.04 (0.87) [0.78]	–	16.82 (0.76)	20.77 (0.94)
NNET ADML	Heat	1.45 (0.66) [0.56]	−0.04	11.94 (0.54)	14.74 (0.67)
NNET ADML	Cold	0.75 (0.86) [0.64]	−0.09	6.17 (0.28)	7.62 (0.35)
NNET ADML	Joint	2.17 (0.86) [0.90]	−0.14	17.90 (0.81)	22.09 (1.00)
Full OLS + NNET ADML	Joint	2.07 (0.99) [0.85]	–	17.06 (0.78)	21.06 (0.96)
Panel C. Stress days, year FE					
Full OLS	Heat	1.70 (0.69) [0.40]	–	13.93 (0.63)	17.20 (0.78)
Full OLS	Cold	0.58 (0.60) [0.37]	–	4.71 (0.21)	5.82 (0.26)
Full OLS	Joint	2.27 (0.98) [0.55]	–	18.73 (0.85)	23.12 (1.05)
Panel D. Stress load, year FE					
Full OLS	Heat	1.76 (0.88) [0.67]	–	14.50 (0.66)	17.90 (0.81)
Full OLS	Cold	0.26 (0.89) [0.70]	–	2.08 (0.09)	2.57 (0.12)
Full OLS	Joint	2.02 (1.28) [1.20]	–	16.62 (0.76)	20.52 (0.93)

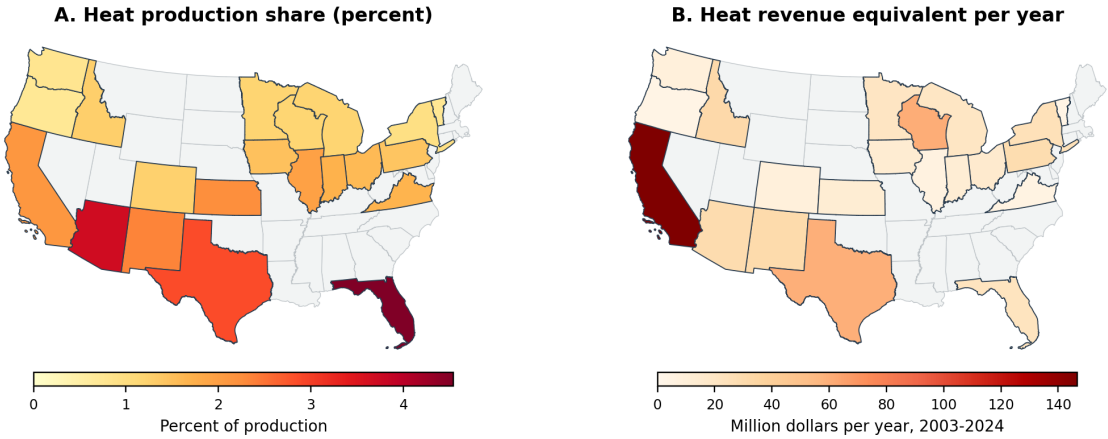
Note: Shares are 100 times log points, with two-way standard errors in parentheses and Driscoll–Kraay in brackets. Plug-in is NNET Plug-in's share from the same fit. Revenue equivalents are for the 48 contiguous states over 2003–2024, in billion dollars, with the average per year in parentheses; the 2024-price column values every year at each state's 2024 all-milk price.

Table A8. Production shares and revenue equivalents by state, 2003–2024.

	Production share (%)		Revenue equivalent (million $)	
State	Heat	Cold	Heat	Heat and cold
Florida	4.54 (1.01)	0.03	498 (22.6)	501 (22.8)
Arizona	3.70 (0.90)	0.10	651 (29.6)	668 (30.3)
Texas	2.87 (0.63)	0.51	1,324 (60.2)	1,561 (71.0)
New Mexico	2.35 (0.52)	0.53	671 (30.5)	824 (37.5)
Kansas	2.25 (0.48)	1.03	279 (12.7)	410 (18.6)
California	2.13 (0.50)	0.50	3,229 (146.8)	4,004 (182.0)
Illinois	1.95 (0.40)	1.20	147 (6.7)	239 (10.9)
Indiana	1.70 (0.35)	1.28	269 (12.2)	474 (21.5)
Virginia	1.68 (0.35)	0.90	125 (5.7)	193 (8.8)
Ohio	1.58 (0.33)	1.21	350 (15.9)	623 (28.3)
Iowa	1.49 (0.31)	1.85	291 (13.2)	659 (30.0)
Pennsylvania	1.43 (0.30)	1.18	639 (29.0)	1,173 (53.3)
Idaho	1.29 (0.31)	1.71	695 (31.6)	1,628 (74.0)
Colorado	1.26 (0.29)	1.60	192 (8.7)	441 (20.0)
Minnesota	1.20 (0.26)	2.24	462 (21.0)	1,338 (60.8)
Michigan	1.19 (0.26)	1.72	453 (20.6)	1,116 (50.7)
Wisconsin	1.18 (0.26)	1.99	1,329 (60.4)	3,609 (164.1)
New York	0.95 (0.22)	1.57	547 (24.9)	1,468 (66.7)
Washington	0.78 (0.21)	1.29	196 (8.9)	521 (23.7)
Vermont	0.75 (0.19)	1.91	82 (3.7)	293 (13.3)
Oregon	0.72 (0.20)	0.78	76 (3.5)	160 (7.3)
21 states	1.68	1.23	12,503 (568.3)	21,903 (995.6)

Note: Each share is the Full OLS stress-day response (year-month FE, lags 0–3) evaluated at the state's own exposure: the production-weighted mean change in log milk per cow when the tail is removed, in 100 times log points. Two-way standard errors for heat, in parentheses, hold the state's exposure fixed. Cold shares are shown without standard errors because the simulation does not support Full OLS intervals for cold. Revenue equivalents apply each state's share to its own production and all-milk price, summed over 2003–2024 in nominal dollars, with the average per year in parentheses; the last column removes both tails. States are ordered by heat share. The last row gives the production-weighted mean shares, which equal the national Full OLS shares, and the summed revenue.

Figure A1. Heat production share and revenue equivalent by state, 2003–2024.



Note: Full OLS heat shares at each estimation state's exposure (Panel A) and their revenue equivalents per year (Panel B), as in Table A8. Grey states are outside the estimation sample.

References

Advani, A., Kitagawa, T., & Słoczyński, T. (2019). Mostly harmless simulations? using Monte Carlo studies for estimator selection. *Journal of Applied Econometrics*, *34*(6), 893–910. https://doi.org/10.1002/jae.2724

Annan, F., & Schlenker, W. (2015). Federal crop insurance and the disincentive to adapt to extreme heat. *American Economic Review*, *105*(5), 262–266. https://doi.org/10.1257/aer.p20151031

Armstrong, D. V. (1994). Heat stress interaction with shade and cooling. *Journal of Dairy Science*, *77*(7), 2044–2050. https://doi.org/10.3168/jds.S0022-0302(94)77149-6

Burke, M., & Emerick, K. (2016). Adaptation to climate change: Evidence from U.S. agriculture. *American Economic Journal: Economic Policy*, *8*(3), 106–140. https://doi.org/10.1257/pol.20130025

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Chernozhukov, V., Newey, W. K., & Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, *90*(3), 967–1027. https://doi.org/10.3982/ECTA18515

Chiang, H. D., Kato, K., Ma, Y., & Sasaki, Y. (2022). Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics*, *40*(3), 1046–1056. https://doi.org/10.1080/07350015.2021.1895815

Choi, E., Carneiro de Souza, V., Dillon, J. A., Kebreab, E., & Mueller, N. D. (2024). Comparative analysis of thermal indices for modeling cold and heat stress in US dairy systems. *Journal of Dairy Science*, *107*(8), 5817–5832. https://doi.org/10.3168/jds.2023-24412

Deschênes, O., & Greenstone, M. (2007). The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather. *American Economic Review*, *97*(1), 354–385. https://doi.org/10.1257/aer.97.1.354

Driscoll, J. C., & Kraay, A. C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, *80*(4), 549–560. https://doi.org/10.1162/003465398557825

Fan, X., & Ma, J. (2024). A spatial econometric analysis of weather effects on milk production. *Earth*, *5*(3), 477–490. https://doi.org/10.3390/earth5030026

Fu, X., Zhang, Y., Zhang, Y. G., Yin, Y. L., Yan, S. C., Zhao, Y. Z., & Shen, W. Z. (2022). Research and application of a new multilevel fuzzy comprehensive evaluation method for cold stress in dairy cows. *Journal of Dairy Science*, *105*(11), 9137–9161. https://doi.org/10.3168/jds.2022-21828

Garcia, A., & Kalscheur, K. (2004). *Dairies and local economic development* (tech. rep. No. Fact Sheet FS 925-G). South Dakota State University Cooperative Extension Service. https://openprairie.sdstate.edu/cgi/viewcontent.cgi?article=1172&context=extension_fact

Gisbert-Queral, M., Henningsen, A., Markussen, B., Niles, M. T., Kebreab, E., Rigden, A. J., & Mueller, N. D. (2021). Climate impacts and adaptation in US dairy systems 1981–2018. *Nature Food*, *2*, 894–901. https://doi.org/10.1038/s43016-021-00372-z

Górska-Warsewicz, H., Rejman, K., Laskowski, W., & Czeczotko, M. (2019). Milk and dairy products and their nutritional contribution to the average polish diet. *Nutrients*, *11*(8), 1771. https://doi.org/10.3390/nu11081771

Heaney, R. P. (2014). How sound is the science behind the dietary recommendations for dairy? *The American Journal of Clinical Nutrition*, *99*(5), 1217S–1222S. https://doi.org/10.3945/ajcn.113.073007

Heck, J. M. L., van Valenberg, H. J. F., Dijkstra, J., & van Hooijdonk, A. C. M. (2009). Seasonal variation in the dutch bovine raw milk composition. *Journal of Dairy Science*, *92*(10), 4745–4755. https://doi.org/10.3168/jds.2009-2146

Hsiang, S. (2016). Climate econometrics. *Annual Review of Resource Economics*, *8*, 43–75. https://doi.org/10.1146/annurev-resource-100815-095343

Hutchins, J., Skidmore, M., & Nolan, D. (2025). Vulnerability of US dairy farms to extreme heat. *Food Policy*, *131*, 102821. https://doi.org/10.1016/j.foodpol.2025.102821

International Dairy Foods Association (IDFA). (2025). *Dairy delivers: 2025 economic impact study of the dairy products industry* (tech. rep.). https://idfa.guerrillaeconomics.net/res/Methodology.pdf

Kadzere, C. T., Murphy, M. R., Silanikove, N., & Maltz, E. (2002). Heat stress in lactating dairy cows: A review. *Livestock Production Science*, *77*(1), 59–91. https://doi.org/10.1016/S0301-6226(01)00330-X

Key, N., Sneeringer, S., & Marquardt, D. (2014). *Climate change, heat stress, and u.s. dairy production* (Economic Research Report No. 175). USDA Economic Research Service.

Klosin, S., & Vilgalys, M. (2025). *Estimating continuous treatment effects in panel data using machine learning with a climate application* [arXiv:2207.08789v3]. https://arxiv.org/abs/2207.08789v3

Lopez, R. A., Laughton, C., Kim, D., & Suh, H. S. (2022). Climate change and productivity of northeastern dairy farms. *Agricultural and Resource Economics Review*, *51*(2), 203–221. https://doi.org/10.1017/age.2022.2

Mauger, G., Bauman, Y., Nennich, T., & Salathé, E. (2015). Impacts of climate change on milk production in the united states. *The Professional Geographer*, *67*(1), 121–131. https://doi.org/10.1080/00330124.2014.921017

Mbuthia, J. M., Eggert, A., & Reinsch, N. (2022). Cooling temperature humidity index-days as a heat load indicator for milk production traits. *Frontiers in Animal Science*, *3*, 946592. https://doi.org/10.3389/fanim.2022.946592

McWhorter, T. M., Sargolzaei, M., Sattler, C. G., Utt, M. D., Tsuruta, S., Misztal, I., & Lourenco, D. (2023). Single-step genomic predictions for heat tolerance of production yields in US holsteins and jerseys. *Journal of Dairy Science*, *106*(11), 7861–7879.

Moore, S. S., Costa, A., Penasa, M., Callegaro, S., & De Marchi, M. (2023). How heat stress conditions affect milk yield, composition, and price in italian holstein herds. *Journal of Dairy Science*, *106*(6), 4042–4058.

Mukherjee, D., Bravo-Ureta, B. E., & De Vries, A. (2013). Dairy productivity and climatic conditions: Econometric evidence from south-eastern united states. *Australian Jour-*

nal of Agricultural and Resource Economics, *57*(1), 123–140. https://doi.org/10.1111/j.1467-8489.2012.00603.x

OECD/FAO. (2022). *OECD-FAO agricultural outlook 2022–2031*. OECD Publishing and Food and Agriculture Organization of the United Nations. https://doi.org/10.1787/f1b0b29c-en

Prieto, J., Lin, Z., Reed, K. F., Wolf, C. A., & Ortiz-Bobea, A. (2026). Milk composition responses amplify economic damages from heat stress. *Environmental Research Letters*, *21*(12), 124029. https://doi.org/10.1088/1748-9326/ae74e6

Ravagnolo, O., Misztal, I., & Hoogenboom, G. (2000). Genetic component of heat stress in dairy cattle, development of heat index function. *Journal of Dairy Science*, *83*(9), 2120–2125. https://doi.org/10.3168/jds.S0022-0302(00)75094-6

Salfer, I. J., Dechow, C. D., & Harvatine, K. J. (2019). Annual rhythms of milk and milk fat and protein production in dairy cattle in the united states. *Journal of Dairy Science*, *102*(1), 742–753. https://doi.org/10.3168/jds.2018-15040

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, *106*(37), 15594–15598. https://doi.org/10.1073/pnas.0906865106

St-Pierre, N. R., Cobanov, B., & Schnitkey, G. (2003). Economic Losses from Heat Stress by US Livestock Industries. *Journal of Dairy Science*, *86*, E52–E77. https://doi.org/10.3168/jds.S0022-0302(03)74040-5

Tack, J., Barkley, A., & Nalley, L. L. (2015). Effect of warming temperatures on US wheat yields. *Proceedings of the National Academy of Sciences*, *112*(22), 6931–6936. https://doi.org/10.1073/pnas.1415181112

USDA Economic Research Service. (2023). Dairy data: Per capita consumption of selected dairy products. https://www.ers.usda.gov/data-products/dairy-data/

USDA National Agricultural Statistics Service (NASS). (n.d.). QuickStats database. https://quickstats.nass.usda.gov/

Vroege, W., Dalhaus, T., Wauters, E., & Finger, R. (2023). Effects of extreme heat on milk quantity and quality. *Agricultural Systems*, *210*, 103731. https://doi.org/10.1016/j.agsy.2023.103731

West, J. W. (2003). Effects of heat-stress on production in dairy cattle. *Journal of Dairy Science*, *86*(6), 2131–2144. https://doi.org/10.3168/jds.S0022-0302(03)73803-X

Zaveri, E., & Lobell, D. B. (2019). The role of irrigation in changing wheat yields and heat sensitivity in india. *Nature Communications*, *10*, 4144. https://doi.org/10.1038/s41467-019-12183-9