Supplementary Information for

Weather-driven US milk yield losses and economic damages revealed by 9 million cows

Author List:

Eunkyoung Choi[1*], Frances V. Davenport[2], Ziyi Lin[3], Ariel Ortiz-Bobea[3,4], Kirstan F. Reed[5], Ermias Kebreab[6], Nathaniel D. Mueller[1,7]

Affiliations:

1. *Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, CO, USA*
2. *Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, CO, USA*
3. *Charles H. Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY, USA*
4. *Jeb E. Brooks School of Public Policy, Cornell University, Ithaca, NY, USA*
5. *KER Consulting, Sault Sainte Marie, MI, USA*
6. *Department of Animal Science, University of California, Davis, Davis, CA, USA*
7. *Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA*

* corresponding authors: kyoung.choi@colostate.edu

Supplementary Methods

Data

Daily weather data were obtained from PRISM[1] and AgERA5[2,3]. PRISM (Parameter-elevation Relationships on Independent Slopes Model) provides high-resolution, bias-corrected, station-based climate fields at 4 km resolution across the US since 1895. We used daily maximum and minimum temperatures (°C), maximum and minimum vapor pressure deficits (kPa), mean dew-point temperature (°C), and precipitation (mm) for 2000–2025.

AgERA5, produced by the Copernicus Climate Change Service at the European Centre for Medium-Range Weather Forecasts (ECWMF), is a bias-corrected version of ERA5 reanalysis tailored for agricultural applications, available daily since 1979 at 0.1° (~ 10 km). We used daily maximum and minimum temperatures (°K), maximum and minimum relative humidity (%), incoming solar radiation flux (Wm^2), and wind speed at 10m (m/s).

PRISM vapor pressure deficits were converted to relative humidity (%) using corresponding maximum and minimum temperatures based on conversion equations[4]. AgERA5 wind speed at 10 m was adjusted to 2 m using conversion equations to simulate the surface-level wind speed[4]. In our analyses, all temperatures are in a unit of °C. However, when it was multiplied to solar radiation, temperatures were converted into Kelvin (°K) to avoid zero-multiplication artifacts. For simplicity, we refer to maximum temperature as “daytime temperature”, minimum temperature as “nighttime temperature”, maximum vapor pressure deficit as “daytime VPD”, minimum vapor pressure deficit as “nighttime VPD”, maximum relative humidity as “nighttime RH”, and minimum relative humidity as “daytime RH”.

Methods

Discovering causal relationship

We built a Directed Acyclic Graph (DAG)[5] to uncover the causal relationships between weather and milk yields based on domain knowledge from animal and atmospheric sciences (**Supplementary Fig. 1**). In this framework, “weather” refers to the three-day average preceding each test-day, consistent with animal science evidence on high yield sensitivity to thermal stress during these periods[6–8]. The DAG defines the total causal effect of weather on milk yields, comprising both direct and indirect effects. The direct effect indicates the impact of three-day average weather on milk yields, while the indirect effect includes changes in milk yields via altered cow’s conditions from disease or feed changes in management. Feed availability and quality are affected by weather, but typically on longer time scales than three days from the test-day, and are strongly moderated by management decisions[9]. From the DAG, five variables were identified as necessary to adjust for the estimation of the total causal effect of weather: weather, seasonal/monthly variability, management-related impacts, geographical locations of county, and cow lactation cycle with parity.

We note that our modeling approach differs from standard predictive machine learning (ML), which aims solely to maximize out-of-sample prediction accuracy. For example, since high milking frequency is more likely to lead to higher milk yields per cow, capturing this relationship in the predictive ML is helpful for out-of-sample prediction skills. However, in causality-aware ML, milking frequency is a parent variable that directly causes the outcome. Including it in the model would block the causal pathway from weather to yield, thereby obscuring the true effect of weather. For this reason, milking frequency was excluded from the model specification.

Hierarchical clustering using a spearman correlation matrix

Tree-based models are generally robust to correlated features. However, strong correlations can still degrade performance or obscure variable importance. To address this, we applied hierarchical clustering on the features' Spearman rank-order correlations (**Supplementary Fig. 4**). A clustering distance threshold of 0.6 (≈ spearman correlation < 0.4) was selected based on visual inspection of the dendrogram. From each cluster, a single representative feature was retained to form candidate weather variable sets. This procedure initially produced >40 candidates, with 20 unique weather combinations once dataset overlap was accounted for. Model performance was then compared between PRISM- and AgERA5-based features. PRISM consistently outperformed AgERA5 for the same weather combinations (e.g., All_day_RH vs. All_day_RH_AgERA5 in **Supplementary Table 1**). Therefore, PRISM was prioritized as the source for most variables, while solar radiation and wind speed were included from AgERA5, as they are not available in PRISM.

Training and test set split by considering the full lactation cycle

Cows can calve at anytime during the year, which complicates estimation of herd-specific management trends and biological lactation cycles if data are split strictly by calendar year. A single lactation cycle can span two calendar years, creating potential data leakage into validation or test sets. To prevent this, we assigned a reference year to each parity based on the distribution of test-day records across calendar years:

i. if > 50% of milk records fall within the calving year, the reference year is the calving year;

ii. if 40–50% of records fall in the calving year and calving occurs before September, the reference year is the calving year;
iii. if 40–50% of records fall in the following year and calving occurs after September, the reference year is the year after calving; and
iv. if < 40% of records fall in the calving year, the reference year is the year after calving.

Using these rules, each lactation cycle was fully contained within a single reference year. The dataset was then partitioned into a training period (2000–2021) and test period (2022-2024).

GridSearch procedures

We performed a GridSearch to identify the combination of weather variables that maximized region-specific average yield residuals. To ensure that the optima reflected realistic rather than rare conditions, candidate ranges were restricted to the 5th–95th percentile weather range combination that provides the highest region-specific average yield residuals. To exclude the optimum weather conditions from rare weather, we constrained it to be between the 5th and 95th percentiles of each weather variable. Ranges for each variable were determined based on Accumulated Local Effects (ALE) plots. We selected 4–5 continuous bins (at 3–percentile intervals) surrounding the peak point, which indicated the highest yield response. When two peaks produced similarly high yields, both ranges were retained. Conversely, if an isolated peak was surrounded by negative bins, it was excluded as likely noise. Gridsearch was then applied across all candidate ranges to evaluate every possible combination of weather variables. Region-specific average yields were computed for each combination, incorporating weights from **equation (4)** in main text to reflect USDA cow population shares and to adjust for imbalances in test-day records across cows.

Calculation of cost function

To select the key weather combinations among 20 candidate feature sets, we used a cost function designed to evaluate the robustness of model performance across both cross-validation (CV) and test (out-of-sample) predictions. Performance was assessed using two loss metrics: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). For each candidate model, RMSE and MAE were first computed within each of the five CV folds as well as on the independent test set. To balance in sample and out of sample performance, we averaged each metric across the five CV folds and then took the mean of the CV average and test-set value to generate a final composite score. This approach weights the test set while ensuring stability across folds. Final RMSE and MAE values for each candidate model are reported in **Supplementary Table 1**. The units of both metrics are log-transformed milk residuals (kg per cow per test-day).

Supplemental Reference

1. PRISM Climate Group. Oregon State University. https://prism.oregonstate.edu/.
2. Copernicus Climate Change Service. Agrometeorological indicators from 1979 up to 2019 derived from reanalysis. https://doi.org/10.24381/CDS.6C68C9BB (2019).
3. Boogaard, H. *et al.* Agrometeorological indicators from 1979 to present derived from reanalysis. https://doi.org/10.24381/cds.6c68c9bb.
4. *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*. (Food and Agriculture Organization of the United Nations, Rome, 1998).
5. Textor, J., Van Der Zander, B., Gilthorpe, M. S., Liśkiewicz, M. & Ellison, G. T. H. Robust causal inference using directed acyclic graphs: the R package ‘dagitty’. *Int. J. Epidemiol.* dyw341 (2017) doi:10.1093/ije/dyw341.
6. West, J. W., Mullinix, B. G. & Bernard, J. K. Effects of Hot, Humid Weather on Milk Temperature, Dry Matter Intake, and Milk Yield of Lactating Dairy Cows. *J. Dairy Sci.* **86**, 232–242 (2003).
7. Bryant, J. R., López-Villalobos ,N., Pryce ,J. E., Holmes ,C. W. & and Johnson, D. L. Quantifying the effect of thermal environment on production traits in three breeds of dairy cattle in New Zealand. *N. Z. J. Agric. Res.* **50**, 327–338 (2007).
8. Hagiya, K. *et al.* Length of lags in responses of milk yield and somatic cell score on test day to heat stress in Holsteins. *Anim. Sci. J. Nihon Chikusan Gakkaiho* **90**, 613–618 (2019).
9. Adams, C., Clark, R. T., Kloptenstein, T. J. & Volesky, D. Matching the Cow with Forage Resources.
10. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

Supplementary Figures

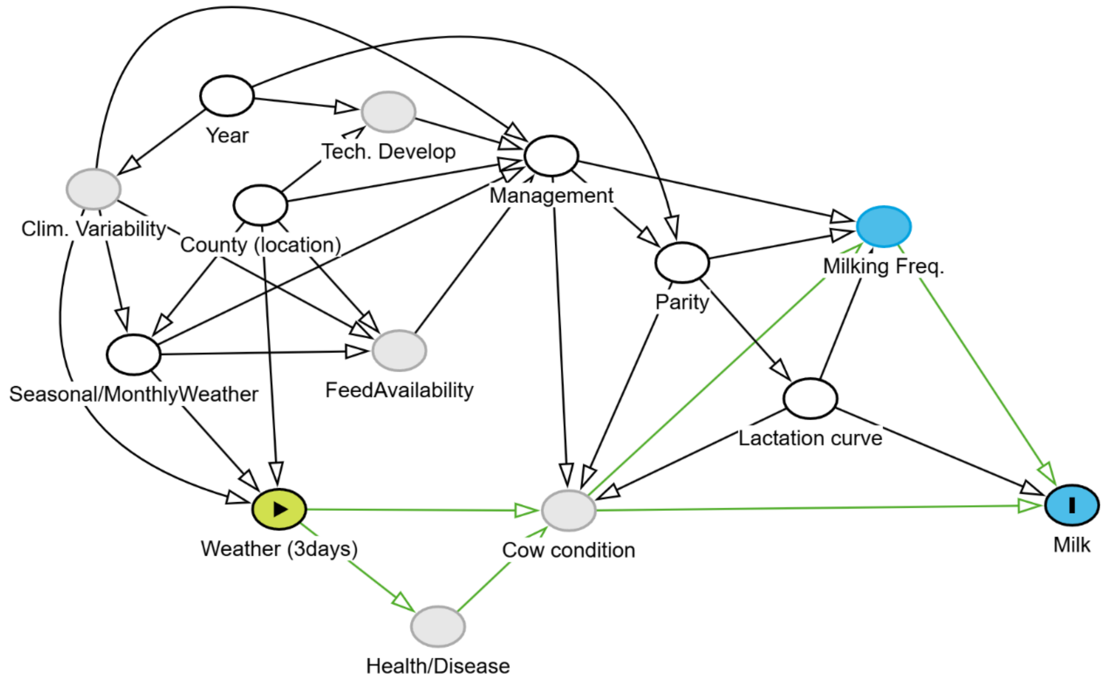

Fig. S1. Directed Acyclic Graph illustrating the causal relationship between weather and milk yields. The solid green oval with a triangle indicates the exposure, which is weather. The solid blue oval with a straight black line presents the target, which is milk yield. The solid blue oval with no line shows a parent variable for the target. Solid white ovals represent features that we can control for, while solid gray ovals represent unobserved features that cannot be controlled for. Each feature is modeled as a function of all features with arrows pointing to it. Green line arrows show the causal paths from the exposure to the target. In our DAG, multiple green line arrows indicate the total effect of weather on milk yields which combines both direct and indirect pathways through health and disease, cow condition, and milking frequency.

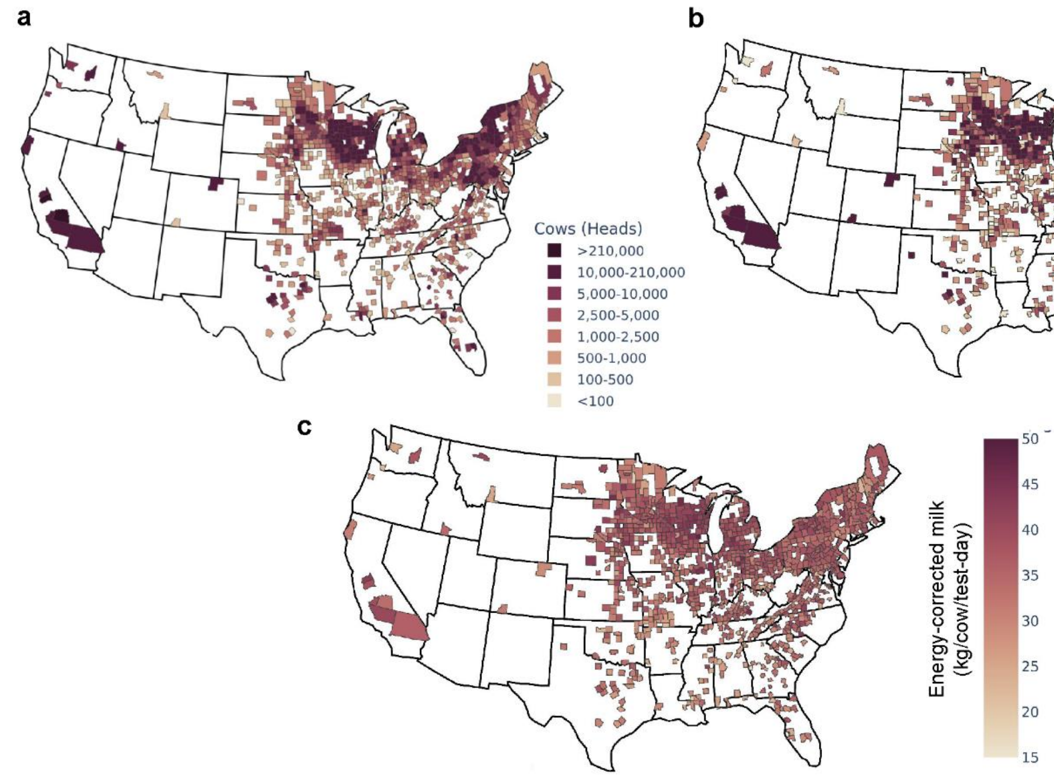

Fig. S2. County-level map of cow populations and milk yields. (a) USDA cow populations averaged across Census years (1997, 2002, 2007, 2012, 2017, and 2022). Counties that are not included in our sample are excluded. (b) Our sample-based (DRMS) cow populations averaged annually across 2000–2024. The legend for (a) and (b) is included between two panels. (c) Energy-corrected milk yields averaged annually across 2000–2024.

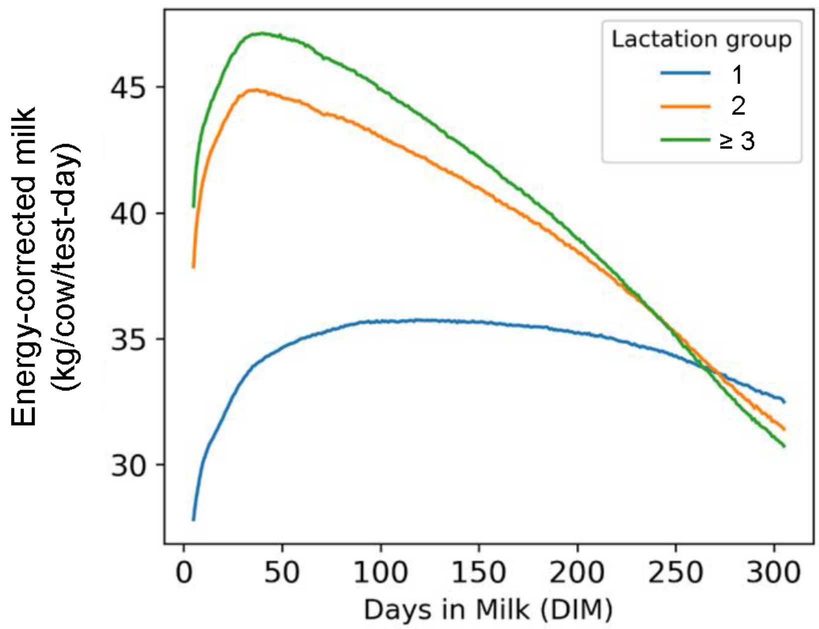

Fig. S3. Cows' lactation cycles across parities. Energy-corrected milk yields were averaged across cows in the US. Lactation group 3 includes all parities greater than three.

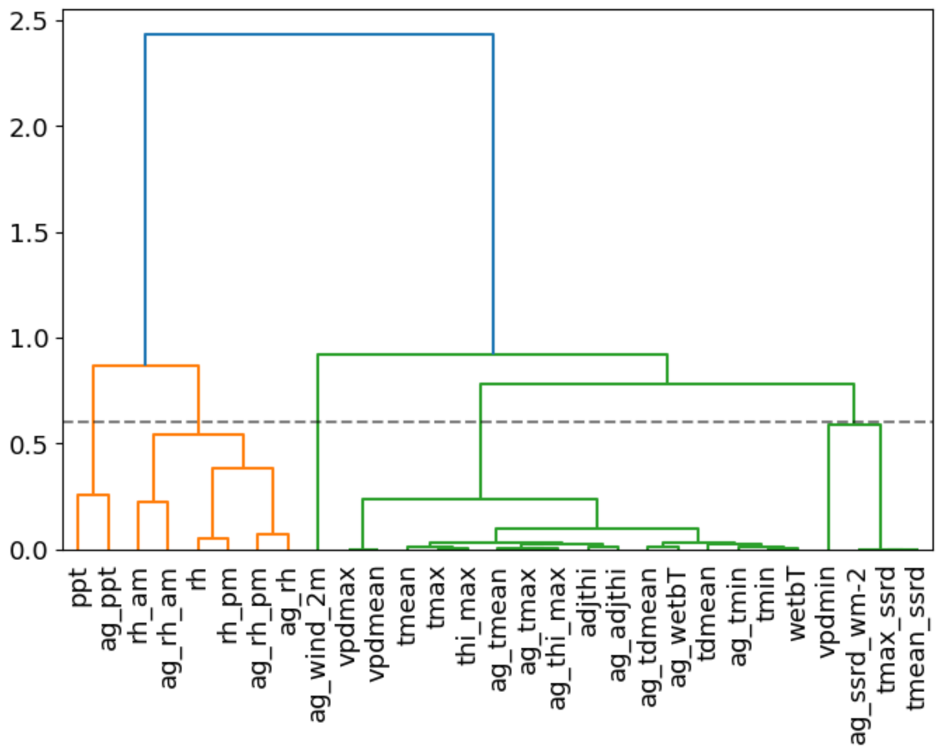

Fig. S4. Hierarchical clustering of weather variables based on a spearman correlation matrix. A distance threshold of 0.6 was applied to define clusters. The prefix "ag_" denotes variables from the AgERA5 dataset, while unlabeled variables correspond to PRISM dataset. Variable abbreviations are as follows: ppt = precipitation, tmax/tmin/tmean = maximum/minimum/mean temperature, tdmean = mean dew-point temperature, wetbT = wet-bulb temperature, vpdmax/vpdmin/vpdmean = maximum/minimum/mean vapor pressure deficits, rh_am/rh_pm = morning/afternoon relative humidity, wind_2m = wind speed at 2m height, ssrd = surface solar radiation downwards, thi_max = maximum Temperature and Humidity Index, and adjthi = adjusted Temperature and Humidity Index.

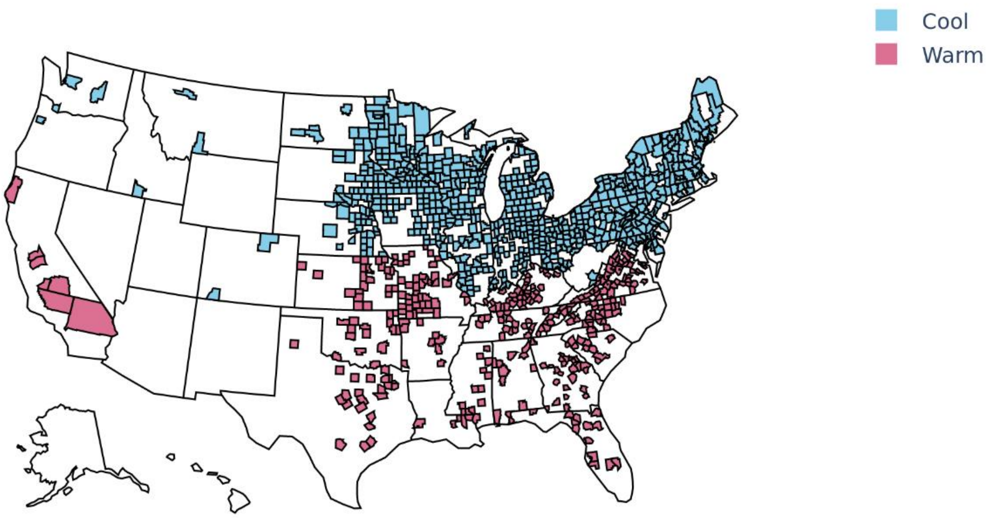

Fig. S5. County-level map of warm and cool regions

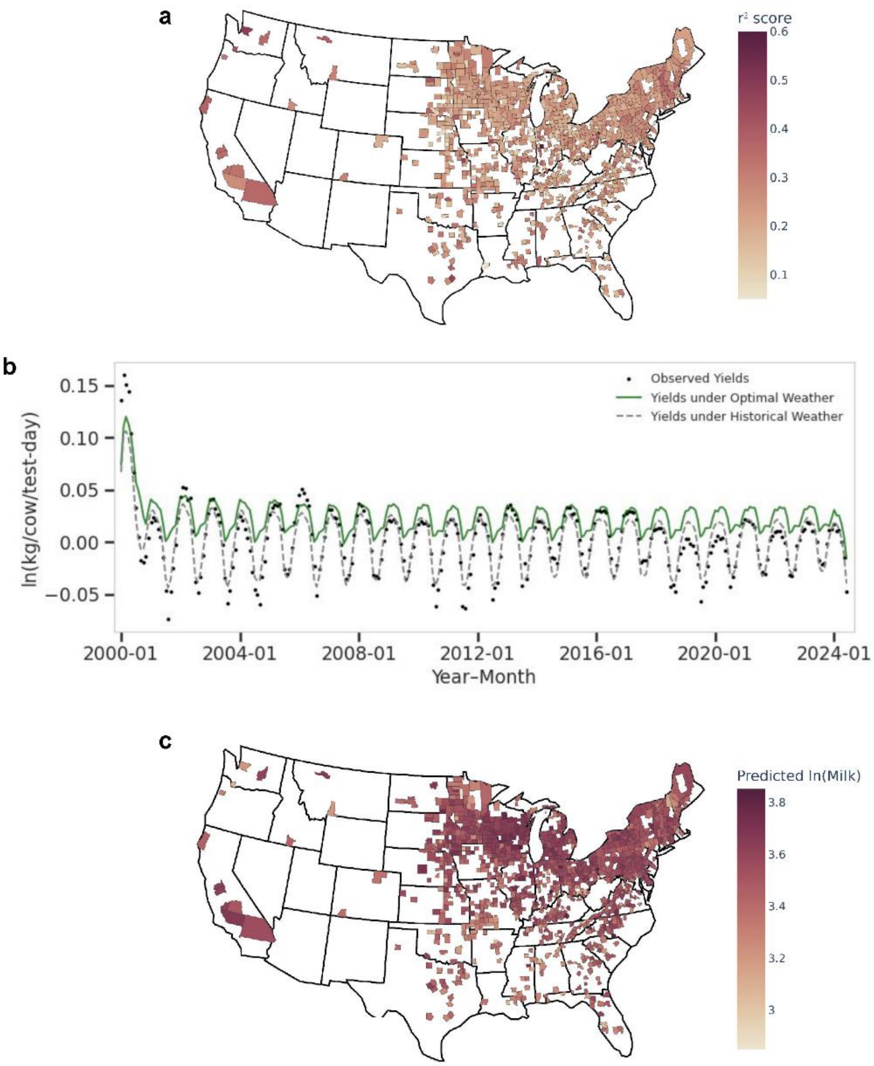

Fig. S6. Spatial and temporal evaluation of model predictions. (a) County-level predictive skill (R^2) of the final XGBoost model. (b) Timeseries of national average milk residual predictions under historical and optimum weather compared with observed residuals. (c) County-level spatial distribution of predicted values from the final XGBoost model, combined with herd-level management trends. Note that seven counties with poor predictive skill (negative R^2) were excluded for yield loss and economic damage analyses.

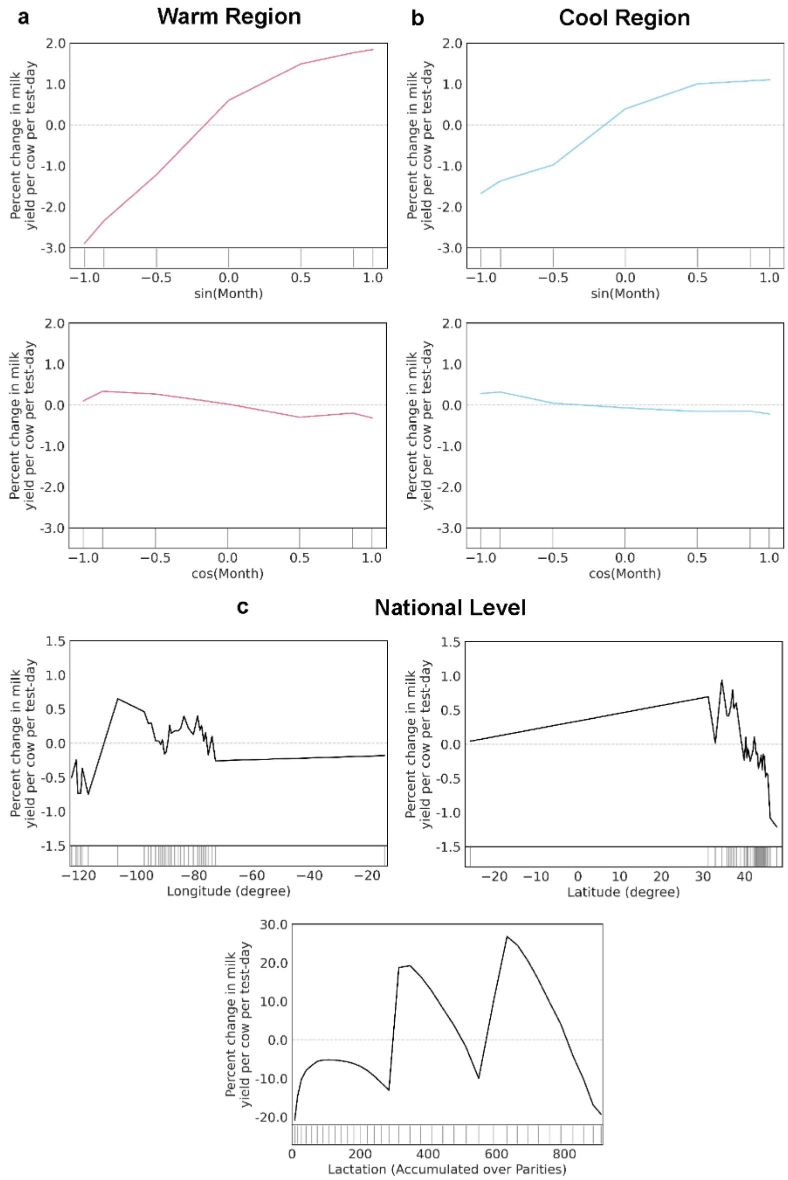

Fig. S7. Average milk yield sensitivity to non-weather variables in (a) warm region, (b) cool region, (c) nationwide. Solid lines show the estimated marginal change (%) in milk yield per cow per test-day to each non-weather variable using Accumulated Local Effects. The pink line indicates the yield sensitivity in warm region, the blue line indicates the yield sensitivity in cool region, and the black line indicates national level yield sensitivity. The rug plot on the bottom of each panel describes the distribution of weighted observations across percentiles (**Methods**).

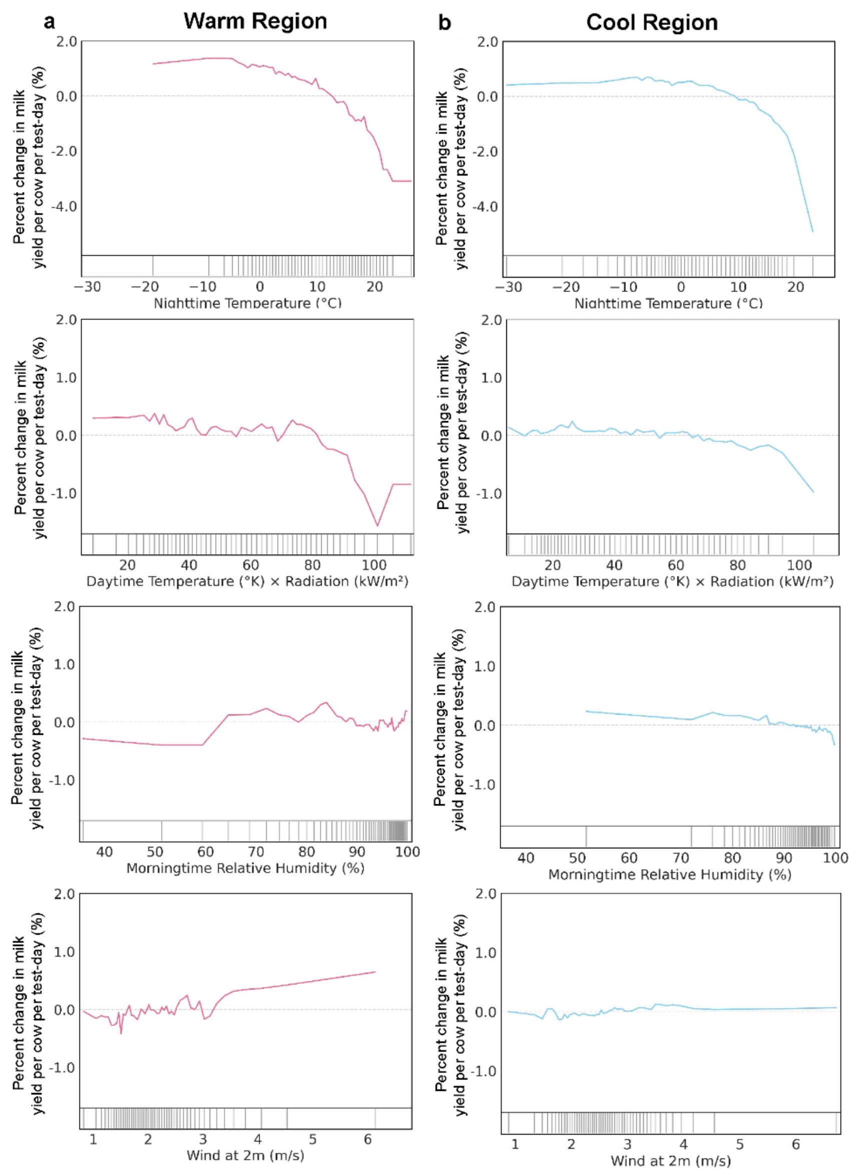

Fig. S8. Aaverage milk yield sensitivity to weather variables in (a) warm and (b) cool regions, 2000–2024. without weighting test-day record imbalance and USDA cow population differences. Solid lines show the estimated marginal change (%) in milk yield per cow per test-day to each weather variable using ALE, in pink for warm region and in blue for cool region. Positive and negative values denote predictions that are above or below the regional mean, respectively. The rug plot on the bottom of each panel describes the distribution of observations across percentiles (**Methods**).

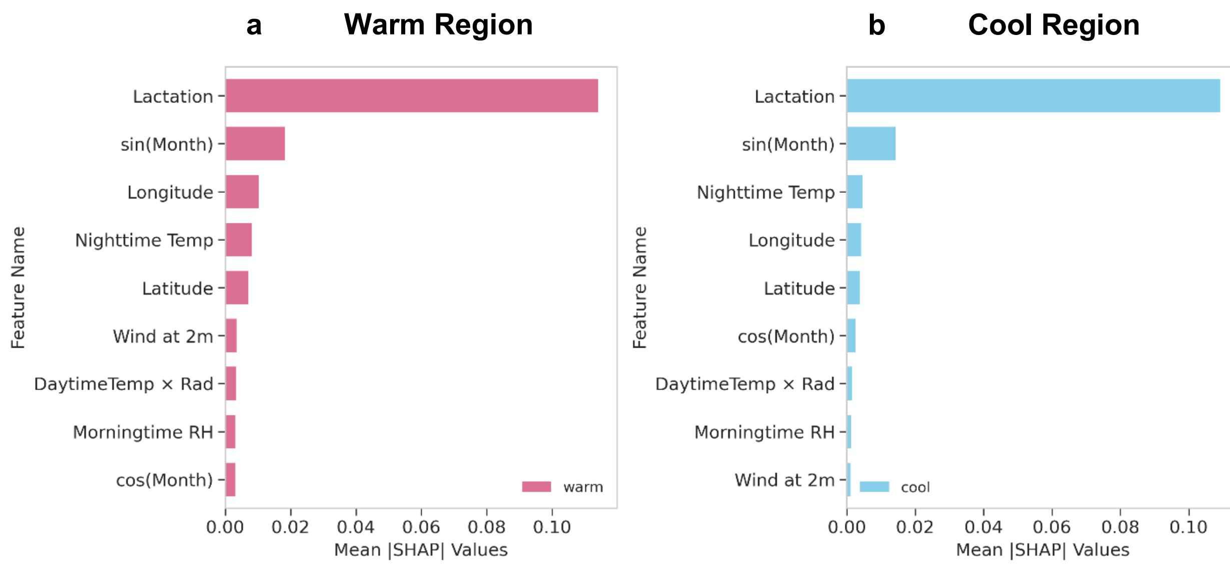

Fig. S9. Feature importance in predicting milk yield changes for (a) warm and (b) cool regions, using SHapley Additive exPlanations[10].

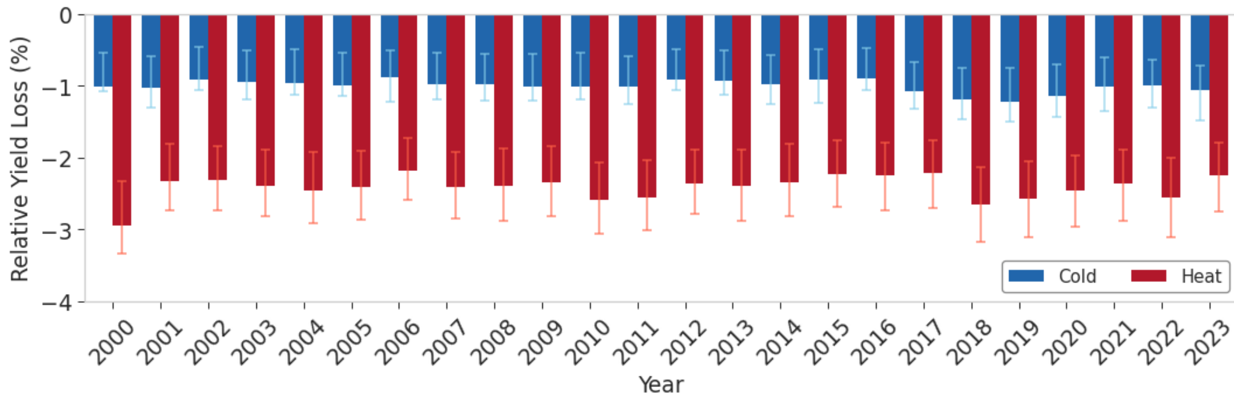

Fig. S10. National-level annual average relative yield loss (%) for heat stress in red and cold stress in blue, 2000–2023.

Supplementary Tables

	Features																				Average Score across CV & Test		
Candidates	Lact-ation	Season	Tavg	Tmin	Tmax	WetbT	Tdew	VPD avg	VPD min	VPD max	RH avg	RH min	RH max	Rad	Tavg x Rad	Tmax x Rad	Wind	PPT	THI max	Adj THI	RMSE	MAE	Final
Baseline	√																				0.248340	0.182854	0.215597
NightRH	√	√		√									√			√	√				0.245525	0.180430	0.212977
NightVPD	√	√		√					√							√	√				0.245527	0.180430	0.212978
DayVPD	√	√		√						√						√	√				0.245524	0.180435	0.212980
Wetbt_Solar	√	√				√										√	√				0.245523	0.180438	0.212981
All_day_RH	√	√		√								√	√			√	√				0.245527	0.180434	0.212981
NightVPD_PPT	√	√		√					√							√	√	√			0.245528	0.180434	0.212981
Wetbt_Night_Solar	√	√		√		√							√			√	√				0.245526	0.180437	0.212982
DayRH	√	√		√								√				√	√	√			0.245529	0.180435	0.212982
Basic2	√	√		√								√			√		√				0.245529	0.180439	0.212984
Night_RHavg	√	√		√							√					√	√				0.245532	0.180437	0.212984
All_day_RH_agera5*	√	√		√								√	√			√	√				0.245532	0.180438	0.212985
DayVPD_PPT	√	√		√						√						√	√	√			0.245530	0.180442	0.212986
DayRH_PPT	√	√		√								√				√	√	√			0.245533	0.180442	0.212988
Basic1	√	√		√	√		√							√			√	√			0.245531	0.180444	0.212988
Basci3	√	√	√								√					√	√				0.245536	0.180444	0.212990
AdjTHImax_Indiv	√	√			√							√		√			√				0.245536	0.180449	0.212992
AdjTHI_Indiv	√	√	√								√			√			√				0.245550	0.180451	0.213000
AdjTHI	√	√																		√	0.245568	0.180452	0.213010
THImax	√	√																	√		0.245581	0.180466	0.213023
THI_Indiv	√	√	√								√										0.245611	0.180489	0.213050
THImax_Indiv	√	√			√							√									0.245611	0.180502	0.213057

*agera5 version of weather combination candidate

Table S1. Error scores of weather combination candidates across cross-validation and test set predictions. In the features, T stands for temperature (°C), WetbT is wet-bult temperature (°C), VPD indicates vapor pressure deficits (kPa), RH denotes relative humidity (%), Rad is downward incoming radiation flux (Wm^2), PPT is precipitation (mm), THI is Temperature Humidity Index, and adjTHI is adjusted THI. Tavg x Rad indicates the interactions of average temperature (°K) and downward incoming solar radiation (kWm^2), and Tmax x Rad denotes the interactions of maximum temperature (°K) and downward incoming solar radiation (kWm^2). A check mark indicates inclusion of a given feature in a candidate combination. All candidates except the baseline additionally included longitude, latitude, and sine- and cosine-transformed month to capture seasonality, as well as lactation cycles across parities. The baseline model included only lactation cycles. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were computed for cross-validation and test-set prediction separately, and their averages were used as the final composite score to select the key weather combination. Units for RMSE and MAE are log-transformed milk residuals (kg/cow/test-day).

	# of Features <= 7	# of Features > 7
Number of Trees	200	250
Max Depth	10	12
Subsample ratio	0.9	0.9
Minium child weight	200	160
Learning rate	0.1	0.05

Table S2. Final hyperparameters from Bayesian optimization. Hyperparameter tuning was initially performed separately for each weather combination candidate. However, final hyperparameters converged to similar values across models. Differences mainly dependent on whether the number of features was fewer than or greater than seven (weather plus control variables including longitude, latitude, sine- and cosine-months, and lactation cycles across parities).