Bias Evaluation and Correction of Heat Index Forecasts Using Climate Observations in California

Amelia Klaus[1,2], Felicia Chiang[2], and Cody Carroll[1,3]

[1]Data Institute, University of San Francisco
[2]Office of Environmental Health Hazard Assessment
[3]Dept. of Mathematics and Statistics, University of San Francisco

Abstract

California's CalHeatScore system relies on National Digital Forecast Database (NDFD) forecasts to produce daily heat-health rankings, making forecast accuracy central to its reliability. In this work, we evaluate systematic biases in NDFD daily maximum heat index forecasts across California using gridMET observations as a baseline over 2015–2024, and we compare two bias correction approaches: Quantile Delta Mapping (QDM) and a residual U-Net. The raw forecasts show a persistent cold bias along the coast and a sharp seasonal flip to a warm bias in the Central Valley during summer, consistent with unresolved coastal dynamics and unrepresented irrigation effects. On a held-out test period (2022–2024), QDM improves distributional alignment, reducing Earth Mover's Distance (EMD) by 37.9%, but worsens mean absolute error by 43.0%, while the U-Net improves both metrics, reducing EMD by 55.2% and consistently lowering daily errors, particularly in topographically complex coastal regions where QDM degrades performance. Extreme heat days outside of summer remain the most difficult to correct for both methods. These results suggest that spatially aware, data-driven corrections offer meaningful advantages over stationary statistical approaches for heat-health applications.

Introduction

California has experienced numerous extreme heat events in recent years, resulting in environmental damage and impacts on public health (Industrial Economics, Incorporated 2024). As average temperatures continue to rise, so does the need for proper assessment of the risks associated with heat extremes. To address this, California's Office of Environmental Health Hazard Assessment (OEHHA) developed CalHeatScore which takes a step towards quantifying potential health consequences due to extreme heat. CalHeatScore is a system that forecasts localized rankings of heat-health risks at the ZIP code level to support California communities in navigating dangerous weather conditions ("CalHeatScore" 2025). As CalHeatScore relies on daily weather forecasts to produce its rankings, understanding and improving the accuracy of these forecasts are important factors in making advancements to the system.

Weather forecasting relies on quantitative representations of the physical processes that govern the climate system. Numerical weather prediction models aim to simulate interactions between the atmosphere, land, and oceans to form estimates of climate variables among spatial and temporal scales (Ghil et al. 2002). While numerical weather prediction models have become more advanced over time, the physical systems these models attempt to represent are incredibly complex, chaotic, dynamical systems, inevitably leading to forecasting errors.

Deficiencies in model performance can be attributed to model shortcomings and model errors. Model shortcomings occur when a model inadequately represents parts of the climate system (e.g., physical violations, missing physical processes, coarse spatial resolution) (Teutschbein and Seibert 2013). Model shortcomings can result in model errors, which can be defined as unsystematic (random) or systematic differences between model simulated conditions and observations (Teutschbein and Seibert 2013). Model errors stem from issues in initial or boundary conditions, physical and numerical formulations, or general model shortcomings (Teutschbein and Seibert 2013). Unsystematic errors are random variations in the simulated output of the model. These originate from the internal variability of climate models as they serve as unpredictable random noise. In comparison, systematic errors, also known as model biases, are consistently inaccurate model simulations in comparison to real-world observations (Teutschbein and Seibert 2013). We can detect patterns within these systematic errors and use these patterns to correct these areas of repeated mistakes, which is an approach also known as bias correction.

The simplest approach to bias correction is the delta-change method. This method applies a constant shift to records using the mean difference between the modeled and observed climate variable of interest (Maraun 2016). This method assumes the model's bias is uniform as it only corrects the signal mean and no other components of the signal's distribution. Because all values are treated equally by this method, it does not correct any sort of variability or extremes. Therefore, this method is unsuitable for applications focused on extreme heat. The delta-change method also makes the assumption of stationarity, so it does not account for temporal shifts in climate variability (Teutschbein and Seibert 2013). The stationarity assumption is the assumption that the relationship between modeled and observed distributions remain constant as time progresses, which may not be valid under future climate signals (Cannon, Sobie, and Murdock 2015).

Quantile Mapping (QM) is another widely used method of bias correction which adjusts modeled climate data so that its cumulative distribution function (CDF) aligns with the historical observations (Teutschbein and Seibert 2012). To implement this, QM constructs a transfer function based on the CDF of observed and modeled data, and then applies this mapping to future model outputs to correct the entire distribution. This approach performs well at reproducing the historical distribution of temperature or precipitation but also falls victim to the stationarity assumption. Despite its ability to perform well for historical biases, the largest distortions resulting from the stationarity assumption occur in the extremes. This is because extreme values fall in the tails of the distribution where observed data is sparse, meaning the transfer function is fit on very few data points in these regions. When future conditions push temperatures beyond the range seen in the historical training period, the correction has little data to rely on and performs poorly precisely where accurate prediction matters most for heat applications (Cannon, Sobie, and Murdock 2015).

Quantile Delta Mapping (QDM) is an advanced form of QM that addresses the inability of basic QM to preserve climate trends that may exist in the modeled output (Thrasher et al. 2012). It explicitly resolves historical correction biases while preserving quantile-specific changes, which are relative for precipitation and absolute for temperature. It first detrends the model output to remove any climate change signal, then applies quantile mapping to align these detrended values with the historical distribution. Finally, the relative changes are reintroduced to preserve them. By removing these changes before bias correcting, the relative changes are preserved and not distorted during the correction (Cannon, Sobie, and Murdock 2015). While QDM may preserve the presence of a climate change signal in model output, it is not able to address the stationarity problem. In preserving the model's projected changes, it still does not account for the fact that a model's biases may differ in the future.

Another approach to bias correction is the use of deep learning models, which can learn complex spatial patterns in forecast errors. The U-Net is a convolutional neural network originally used for biomedical image segmentation (Ronneberger, Fischer, and Brox 2015) and has been applied to a variety of domains including the bias correction of numerical weather prediction (NOAA Environmental Modeling Center 2025). Unlike QDM, which corrects each grid cell independently, the U-Net's convolutional structure allows each correction to depend on the spatial context, which is useful in areas such as the California coast where forecast errors vary significantly.

To advance the accuracy of CalHeatScore's daily heat-health forecasts, this project first conducts an evaluation of the biases in the weather forecasts that CalHeatScore is based on. Using historical climate observations as a baseline, historical forecast records that correspond to the historical observations can be used to quantify systematic errors. Following the bias evaluation, this project applies two of the bias correction approaches described above to correct the forecasts. First, QDM is used to adjust each forecast value based on where it falls in the historical quantile mapping between the forecast and observed distributions. Second, the residual U-Net approach is used to learn and apply spatially varying corrections across the forecast grid. With these two methods, this project evaluates whether a data-driven spatial model can outperform a traditional statistical approach, particularly in regions with complex terrain and coastal dynamics where stationary corrections tend to struggle.

Data and Methods

CalHeatScore's underlying ranking approach uses the daily maximum heat index from historical climate observations from gridMET (Abatzoglou 2013). In addition, CalHeatScore uses the daily maximum heat index from National Weather Service (NWS) National Digital Forecast Database (NDFD) (National Weather Service 2026b) forecasts to produce its daily heat-health ranks. Therefore, this project uses the heat index from these datasets to mirror the approaches taken by the current CalHeatScore system (Office of Environmental Health Hazard Assessment 2026).

gridMET

gridMET is a spatial dataset with a grid resolution of 4 km, or 1/24° covering the contiguous US (CONUS), with dates ranging from 1979 through the present. To match the current CalHeatScore methodology, we find the daily maximum heat index by using daily maximum air temperature and minimum relative humidity from the dataset. The heat index, HI, is computed using the approach described by the National Weather Service (National Weather Service 2026a), which first applies a simplified formula:

$$HI_{\text{simple}} = 0.5 \times (T + 61.0 + (T - 68.0) \times 1.2 + R \times 0.094) \tag{1}$$

where T is air temperature in degrees Fahrenheit and R is relative humidity as a percentage. If the average of HI_{simple} and T is less than 80°F, this simplified value is used as the final heat index. Otherwise, the full Rothfusz regression is applied:

$$\begin{aligned} HI = & -42.379 + 2.04901523\,T + 10.14333127\,R - 0.22475541\,TR \\ & - 6.83783 \times 10^{-3} T^2 - 5.481717 \times 10^{-2} R^2 \\ & + 1.22874 \times 10^{-3} T^2 R + 8.5282 \times 10^{-4} T R^2 \\ & - 1.99 \times 10^{-6} T^2 R^2 \end{aligned} \tag{2}$$

Two adjustments are then applied to the Rothfusz result under specific conditions. When $80°\text{F} \leq T \leq 112°\text{F}$ and $R < 13\%$, a low-humidity adjustment is subtracted:

$$\text{Adjustment}_{\text{low}} = \frac{13 - R}{4} \times \sqrt{\frac{17 - |T - 95|}{17}} \tag{3}$$

When $80°\text{F} \leq T \leq 87°\text{F}$ and $R > 85\%$, a high-humidity adjustment is added:

$$\text{Adjustment}_{\text{high}} = \frac{R - 85}{10} \times \frac{87 - T}{5} \tag{4}$$

These adjustments account for the fact that high humidity amplifies the perceived heat while low humidity reduces it relative to what the base regression predicts.

Since we intend to compare gridMET with NDFD and gridMET's native resolution is coarser than NDFD's, we adjust the resolution of gridMET by splitting each gridMET cell into four. This doubles the resolution in both latitude and longitude so the new resolution is roughly 2 km or 1/48°. It is important to note that this resolution change was done through nearest-neighbor interpolation, meaning that each new smaller cell is a copy of the parent cell to preserve the statistical properties of gridMET's distribution.

NDFD

The National Weather Service (NWS) National Digital Forecast Database (NDFD) provides gridded forecasts of a variety of weather elements based on output from an ensemble of numerical weather prediction models. Similar to gridMET, NDFD's datasets are spatial, covering the contiguous US with multiple variables, but the preprocessing of NDFD was more involved as it required a multi-staged pipeline with data acquisition, reduction, and regridding.

The raw data comes from NDFD in which two CONUS variables are run through parallel pipelines. Air temperature forecasts, YEU, and relative humidity forecasts, YRU, are extracted from CONUS level files. The acquisition process first scrapes the live link provided by NOAA to collect URLs for each daily folder. From there, it filters a target month and scrapes the URL to download and unpack each month into a target folder, using disk storage. The biggest challenge at this stage is the volume of data. A single month of one variable takes up around 90 GB, so pulling any more than one month at a time is infeasible. To work around this, months are processed one at a time to maintain disk storage. The NDFD resolution is coarser prior to 2015, which makes the data incompatible with more recent data. We omitted 2025 due to missing data in recent archives. Therefore, our study period is 2015–2024, and we conduct a preliminary analysis on a 3-year subset (2016–2018).

While NDFD has multiple forecast runs for each day, we use one representative run to represent the forecast for that day. The forecast run should be taken on that calendar day, with valid times beginning before 6:00 AM Mountain Time (5:00 AM Pacific Time) to match CalHeatScore operations as of April 2026. All timestamps are first converted from UTC to Mountain Time, and then the files are filtered to find the representative forecast run for the day.

After one representative run has been chosen for each day, we calculate the heat index from air temperature and relative humidity using the Rothfusz regression equation described in Equation (2). Then we collapse the valid time dimension by finding the daily maximum heat index at each grid cell for that calendar day. Finally, NDFD must be regridded to align with gridMET so that bias correction techniques can be properly applied. Using bilinear interpolation, NDFD's data is interpolated to match gridMET's grid. At this stage, both sources are on the same scale both spatially and temporally, so they are eligible for comparison.

Bias Correction

Quantile Delta Mapping

Following the preprocessing of both sources, the next step is the first bias correction approach: Quantile Delta Mapping (QDM). Initially, the model is trained on three years of data (2016–2018) by fitting separate quantile mappings for each month. This allows for each month to be bias adjusted separately, which is important as biases can vary by month. After this training, a set of quantile transfer functions are applied to adjust the raw forecast values.

One parameter that was further explored was the number of quantiles used to represent each distribution. Using more quantiles results in more precise corrections, but risks overcorrecting the forecast to match the observations identically. Using too few quantiles did not adapt the shape of the observed distribution. Several values were tested, but ultimately 20 quantiles struck the right balance, as shown in Table 1.

Residual U-Net

To introduce a more spatially and temporally complex approach to bias correction, a Residual U-Net was used. The U-Net architecture used in this project follows the design described in (Hamill et al. 2023). The network consists of two main parts: an encoder that compresses the input and a decoder that reconstructs it back to the original resolution. The encoder reduces the spatial resolution of the input while extracting features from the forecast field. The decoder then upsamples back to the original grid. Skip connections link matching levels of the encoder and decoder, passing spatial details such as coastlines directly to the decoder so that they are not lost during compression. The full network contains $\sim$31 million trainable parameters.

Rather than predicting the corrected field directly, the network predicts an additive correction that is added on top of the raw forecast. This means that the network learns the difference between the forecast and the observations rather than the full heat index field.

Results

Bias Evaluation: 2016–2018

Before exploring bias correction techniques, it is important to analyze the raw errors to establish a baseline for comparison. The raw errors between forecast (NDFD) and observed (gridMET) heat index values were characterized to identify both spatial and temporal bias patterns. Due to data processing constraints, preliminary analysis was performed for the years of 2016–2018.

The first metric, shown in Figure 1, is the monthly average error, computed as the average of the daily difference between the forecast and observed values at each grid cell across the 2016–2018 period. Negative values (blue) are areas of cold bias, meaning the forecast predicts cooler temperatures than observed. Positive values (red) indicate a warm bias in which the forecast predicts warmer temperatures than observed.

Examining these errors spatially, clear regional patterns emerge. Along the coast, there is a persistent cold bias throughout the entire year, but it is most intense from April through July, as this blue signal is the strongest. This bias weakens around August through November but never fully disappears. This persistent coastal cold bias may be attributed to a few factors. First, there is a resolution mismatch between the two datasets. Even after rescaling gridMET to a finer grid, the temperature values in those new cells are still derived from the original coarser resolution data, meaning coastal grid cells that sit near the land and ocean boundary may be misclassified when they are influenced by cooler ocean air temperatures. Second, forecast models generally struggle to capture temperature behavior that exists along the California coast due to the complexity of coastal topography, making this region difficult to predict accurately (Dorman et al. 2001).

The Central Valley shows a different pattern as the year begins with a cold bias from January to May, but a very sharp flip occurs around June when a strong warm bias begins, which lasts through October. This abrupt seasonal reversal is likely driven by irrigation activity in the Central Valley, which has a well-documented cooling effect on surface temperatures (Lobell et al. 2008). The underlying models used to generate NDFD forecasts do not fully account for this, causing the forecast to run systematically warmer than what is actually observed on the ground once irrigation ramps up in the summer months (Qian et al. 2020).

Separately, Earth Mover's Distance (EMD) captures differences in the full shapes of the forecast

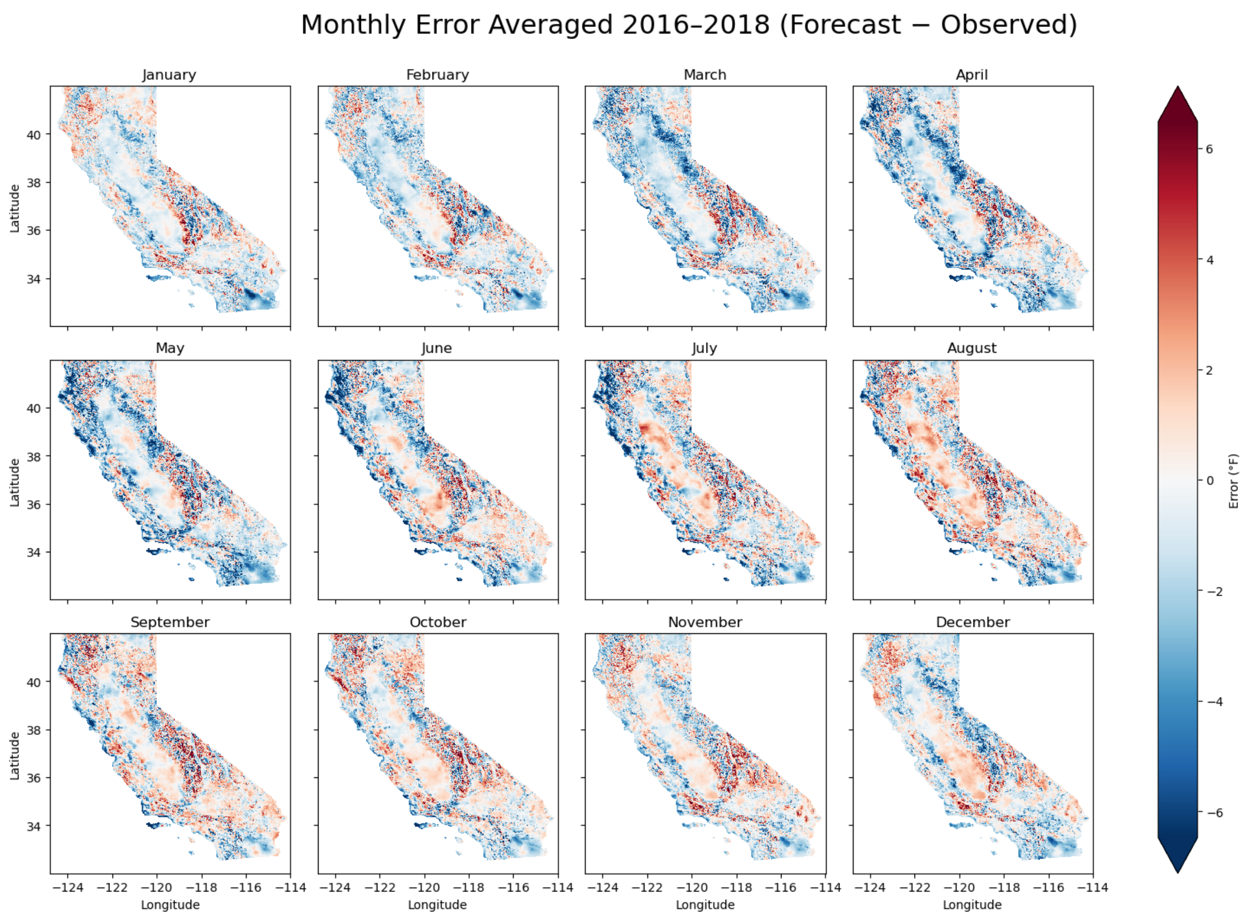

Figure 1: Spatial distribution of monthly mean raw forecast errors across California for the 2016–2018 period, computed as the difference between NDFD forecast and gridMET observed heat index values (Forecast − Observed).

and observation distributions. EMD quantifies the minimal cost that goes into the transformation from one distribution into the other (Rubner, Tomasi, and Guibas 1998). A higher EMD indicates that the two distributions differ more fundamentally, not just on daily inaccuracies.

EMD is computed monthly by pooling all daily heat index values for each month across the three-year window, yielding roughly 90 data points per grid cell. From these, separate forecast and observed distributions are constructed and the distance between them is calculated at each grid cell, producing the spatial maps shown in Figure 2. Looking at our figure, we see larger EMD values surrounding the Central Valley from March through June, where we observed a cold bias in the raw errors. Higher values also appear consistently along the coast, indicating a distributional mismatch along topographically complex areas.

These spatial patterns highlight that forecast bias in California varies substantially by region and shifts across seasons, motivating the use of temporally-stratified correction methods.

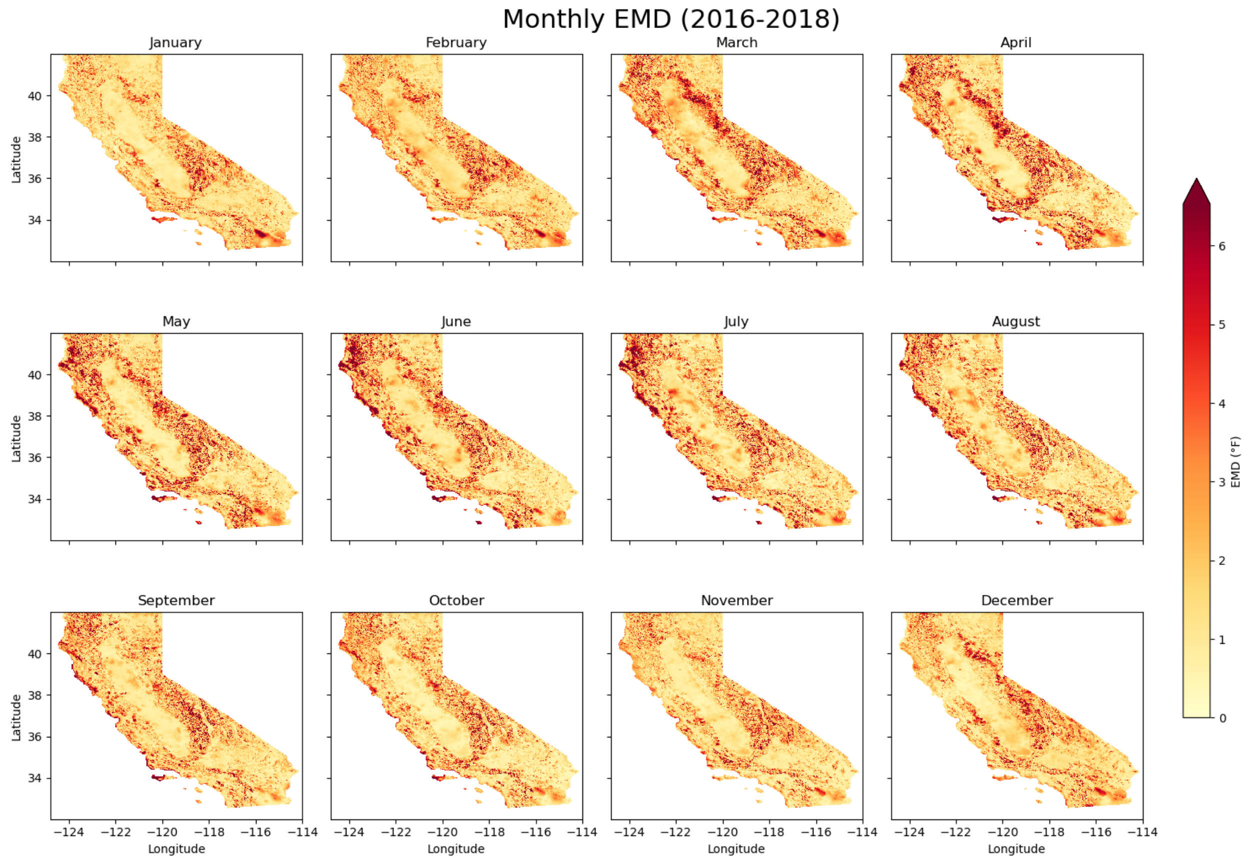

Figure 2: Spatial distribution of Earth Mover's Distance (EMD) across California throughout the 2016–2018 time period.

Initial QDM Analysis: 2016–2018

The first evaluation of QDM evaluates its performance on the same 2016–2018 period it was trained on. While this assessment does not speak to how well the correction generalizes to unseen data, it is a first step to confirm the learned quantile mappings are reducing the raw bias. Figure 3 shows the seasonal mean error (forecast minus observed) before and after applying QDM in all four seasons. The improvement can be seen simply from the change in visual magnitude.

Violin Error Analysis

Violin plots provide another way of visualizing the full error distribution. In Figure 4, the blue violins represent the raw errors and the orange violins represent the QDM-corrected errors.

A consistent pattern emerges across nearly all months: the corrected violins are wider around zero and have their median pulled closer to the zero line, indicating that QDM is successfully reducing both mean bias and the spread of extreme errors. In general, the raw error distributions do not show a strong preference for cool or warm bias, and the overall range of errors is not large to begin with.

Since the ultimate motivation of this project is improving forecast accuracy during dangerous heat

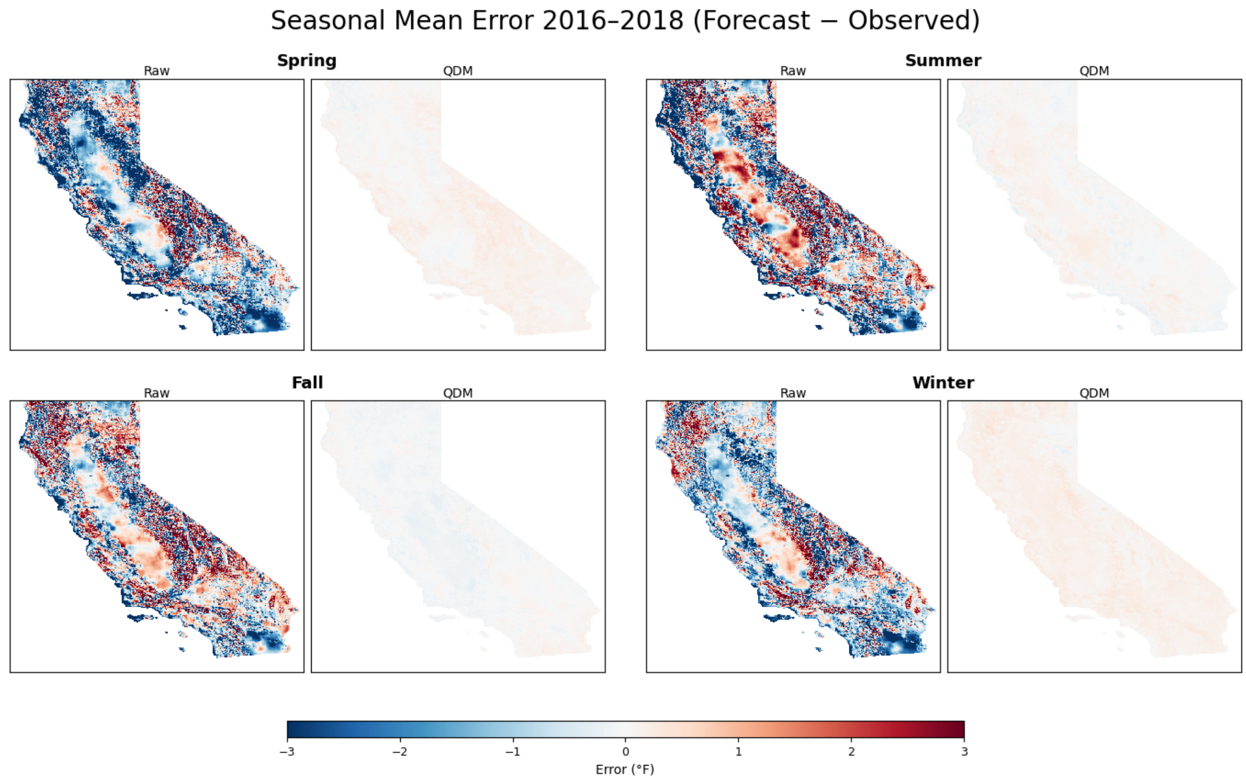

Figure 3: Spatial distribution of seasonal error across California before and after Quantile Delta Mapping (QDM) was applied for the 2016–2018 time period.

events, we specifically evaluate QDM's performance on extreme heat days. The 95th percentile of gridMET observations was computed per grid cell, and this threshold was used as a mask to filter both the raw and corrected errors to only those times and locations where extreme heat was observed. The results for the 2016–2018 training period are shown in Figure 5.

In the warmer months of June through October, the raw errors are already reasonably well-centered around zero and fairly symmetric, and QDM tightens these distributions further while keeping the median near zero, a strong result for the warmest months. The most challenging results emerge in winter and spring. December shows a notably misshapen distribution with compressed tails, skewed heavily in the negative direction indicating the forecast is running cooler than the observations. While QDM reduces the tails somewhat and pulls the median closer to zero, a negative skew remains in the corrected distribution. April and May show raw error tails extending as far as -25°F on extreme days, indicating that when heat events occur in spring, the forecast can substantially under-predict them. QDM shifts the median closer to zero in these months, but the extreme tails in April are not fully eliminated, suggesting that residual errors on hot spring days remain a limitation of the current correction approach. This pattern is consistent with the difficulty the forecast has capturing heat early in the season. It also points to the need for more flexible correction methods that can better handle these extreme values.

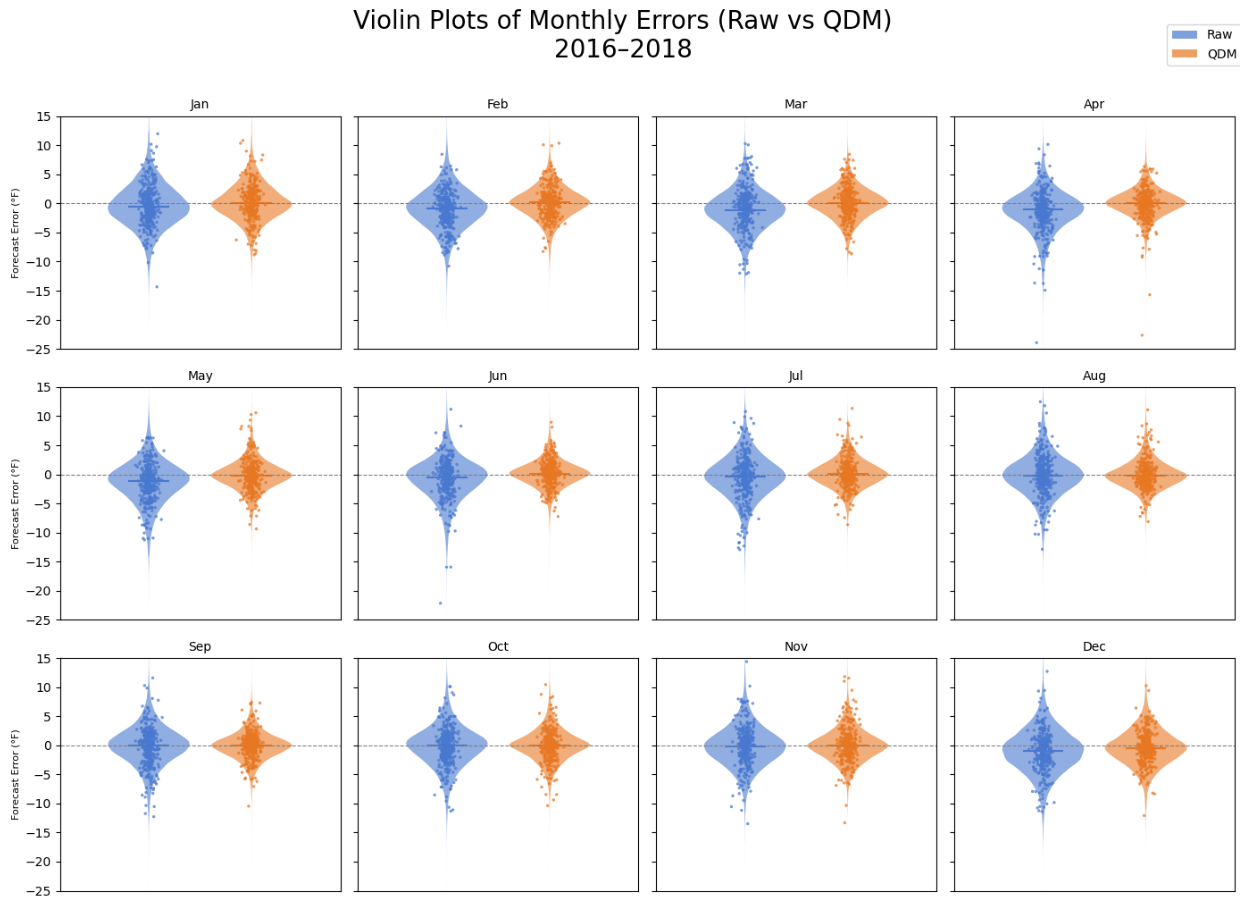

Figure 4: Violin plots of full error distribution throughout the 2016–2018 time period comparing raw errors to QDM.

Comparison of QDM and U-Net

Data Splitting

Following the preliminary analysis of 2016–2018, the pipeline was extended to the full 10-year period (2015–2024). The dataset was divided into training (2015–2020), validation (2021), and testing (2022–2024) sets to use for both QDM and the U-Net approach, to ensure fair comparison across methods. For QDM specifically, the training period was used to learn the quantile mappings, and the validation set was used to select the number of quantiles. A range of values were evaluated against the validation mean absolute error (MAE), shown in Table 1. Interestingly, the differences across quantile counts were negligible with MAE values ranging from approximately 2.26 to 2.28°F with no meaningful trend. To further examine this, the performance of extreme choices (5 vs. 50 quantiles) was evaluated on the test set and again showed no significant difference. This suggests that for this dataset, the number of quantiles does not meaningfully impact the quality of the correction, and the mapping is robust to this parameter choice. For consistency with the earlier preliminary analysis, 20 quantiles was carried forward. The learned mappings from the full training period were then applied across the entire 10-year period.

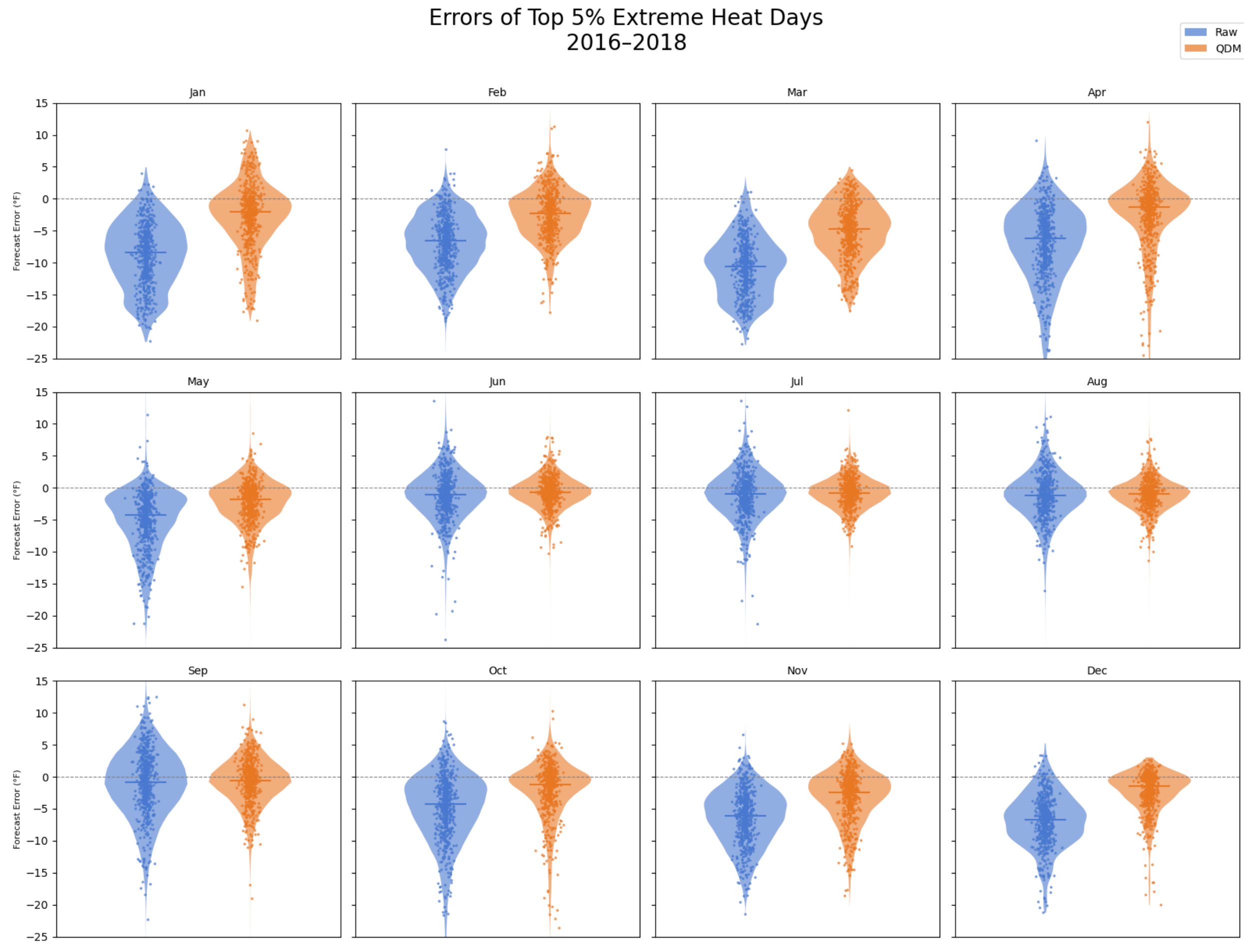

Figure 5: Violin plots of error distribution for 95th percentile of extreme heat days throughout the 2016–2018 time period comparing raw errors to QDM.

EMD Findings

Analysis will now shift to the testing period of 2022–2024 with data that was not used to train or validate either method. QDM's constant mappings learned from the training period were applied directly to this period, and the U-Net was used to generate inferences on the same data.

The first metric used for comparison is Earth Mover's Distance (EMD). Figure 6 shows the seasonal EMD for the raw forecast, QDM, and U-Net across the testing period. Both correction methods show a more noticeable improvement over the raw forecast in EMD than was seen in MAE. QDM improved EMD by 37.9% and the U-Net improved EMD by 55.2% relative to the raw forecast.

MAE Findings

Mean absolute error (MAE) measures the average magnitude of the difference between forecast and observed values at each grid cell, without accounting for the direction of the bias (Willmott and Matsuura 2005). For this project, MAE is an important metric as it directly reflects how far off a heat index prediction is on any given day, which is essential for day-to-day assessment of

Table 1: Validation MAE by Number of Quantiles

N Quantiles	Validation MAE (°F)
3	2.2615
4	2.2656
5	2.2694
2	2.2696
10	2.2752
15	2.2793
20	2.2810
25	2.2819
30	2.2845

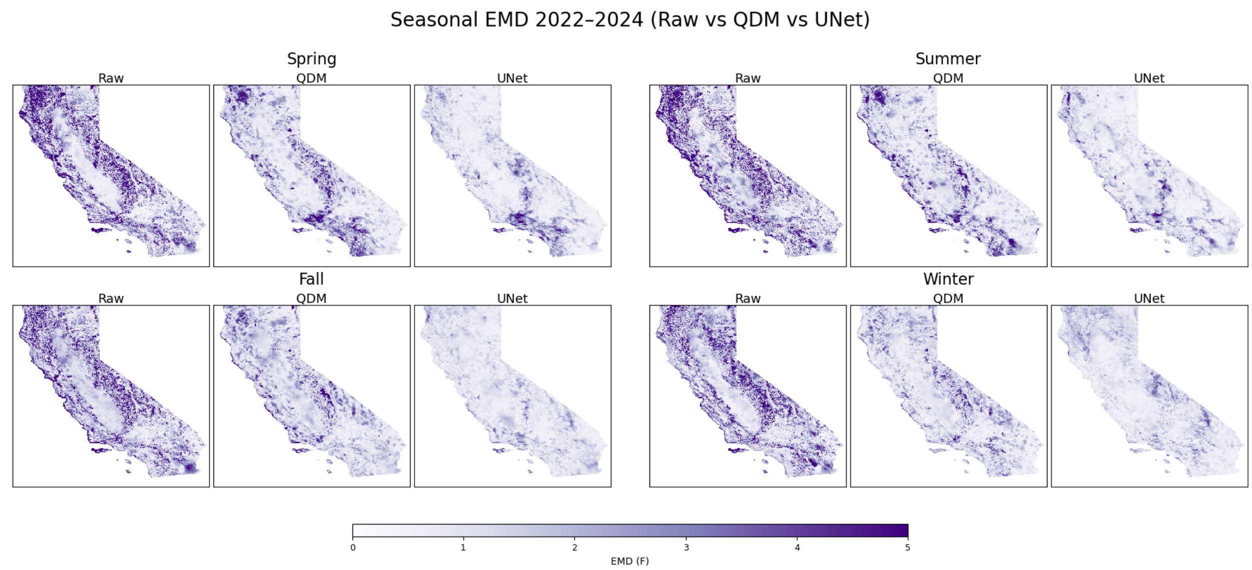

Figure 6: Seasonal distribution of Earth Mover's Distance (EMD) across California throughout 2022–2024 comparing bias correction approaches.

heat-health risk. Shifting from EMD to MAE, Figure 7 shows the spatial MAE across all four seasons for the raw forecast, QDM, and U-Net outputs.

Looking at the differences between the two methods, the U-Net consistently produces the lowest MAE across all seasons. The most notable residual error patches appear in winter for all three fields (raw, QDM, and U-Net) and are not fully addressed by either correction method. Along the coast, QDM reduces but does not fully eliminate the elevated errors, while the U-Net largely dissolves this coastal signal, pointing to the U-Net's ability to capture spatially complex error structure that grid cell-level statistical approach cannot.

Despite improving distributional alignment as seen in the EMD results, QDM actually worsened overall MAE by 43.0% relative to the raw forecast on the test set. While QDM is successfully re-shaping the distribution of predicted heat index values to better match observations, the day-to-day

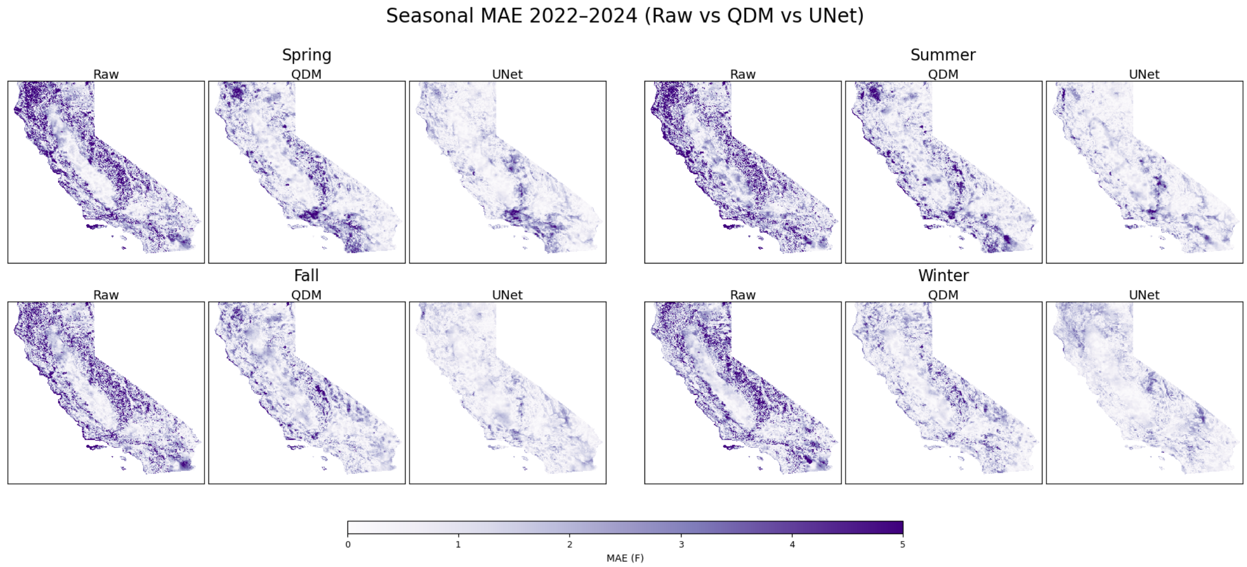

Figure 7: Seasonal distribution of Mean Absolute Error (MAE) across California throughout 2022–2024 comparing bias correction approaches.

predictions are not more accurate. This is consistent with how QDM works by design, as it corrects distributional quantiles rather than optimizing for pointwise accuracy, meaning improvements in distributional shape do not necessarily translate to improvements in daily error. In some regions, the correction overshoots, introducing new error rather than reducing it.

The U-Net, by contrast, performs well on both metrics, reducing both the day-to-day errors and the distributional mismatch. For a heat-health application where daily forecast accuracy matters, this suggests the U-Net is learning a more complete correction.

Grid Cell Analysis

To quantify performance at specific locations across California, Table 2 summarizes the raw MAE, raw EMD, and the percent change in each metric after QDM and U-Net correction for a set of representative grid cells spanning different climate regions.

Table 2: Raw MAE, Raw EMD, and percent change after correction for representative grid cells across California. Negative percentages indicate worsened performance.

Location	Raw MAE (°F)	Raw EMD (°F)	QDM MAE %	UNet MAE %	QDM EMD %	UNet EMD %
San Diego	1.97	0.80	−3.8	2.6	16.5	17.6
San Francisco	2.51	0.51	0.6	6.7	34.9	−12.7
Fresno	2.05	1.08	3.4	13.0	29.7	55.7
Redding	2.42	1.35	0.4	10.9	13.4	32.6
Monterey	2.65	0.71	−9.9	3.6	−65.5	0.9
Eureka	2.58	1.05	−4.1	14.6	35.3	23.2
Sierra Nevada Foothills	2.19	1.49	−0.5	10.8	−2.2	31.7
Lake Tahoe	4.75	4.28	51.2	52.1	86.9	71.7
Mojave Desert	3.17	2.82	6.0	13.7	10.3	16.7
North Coast Ranges	3.83	3.67	42.6	45.0	65.2	67.7
Kern County	3.67	3.33	41.6	42.8	77.0	69.3
Mammoth Lakes East	3.31	2.64	−4.3	2.5	−9.9	−5.6

The U-Net consistently outperforms QDM on MAE across nearly all locations, but the coastal grid cells tell the most striking story. In Monterey, QDM actually worsens MAE by nearly 10% relative to the raw forecast, while the U-Net achieves a modest 3.6% improvement. Eureka shows the same pattern, with QDM degrading MAE by 4.1% while the U-Net improves it by 14.6%. This coastal degradation under QDM is consistent with the broader spatial MAE findings and reinforces that the stationary quantile mappings learned during training do not generalize well to the complex temperature dynamics along the coast. The EMD results add another layer to this picture. Monterey stands out with a -65.5% change in EMD under QDM, meaning the corrected distribution is actually further from the observed distribution than the raw forecast was. This is a failure case where QDM is overcorrecting the distributional shape at this location. The U-Net brings EMD nearly back to neutral at 0.9%, which is still an improvement compared to QDM. Inland locations like Fresno and Redding show positive results for both methods across both metrics.

Lake Tahoe, North Coast Ranges, and Kern County show some of the largest MAE improvements across both methods, each reducing error by over 40% in these locations. This suggests that areas with larger raw errors may have more room for correction gains, and both methods are able to work with this. Mammoth Lakes East is a notable exception where both methods show little to no improvement across either metric, suggesting this high-elevation region presents unique challenges that neither approach fully addresses.

CDF Analysis

To get a closer look at coastal behavior, Figure 8 shows the monthly CDFs at a representative Monterey grid cell, comparing observations against the raw forecast, QDM correction, and U-Net predictions across the full test period.

The raw forecast distribution sits consistently to the left of the observed curve across nearly all months, reflecting the persistent coastal cold bias identified in the error maps. The gap is widest during spring and early summer, consistent with the poor MAE and EMD performance seen at this location in Table 2 for QDM. Rather than pulling the distribution rightward toward the observations, QDM overcorrects or shifts in the wrong direction in several months. This is most visible in May and June, where the QDM curve shifts noticeably away from both the observed and raw forecast curves. This behavior directly explains the -65.5% EMD result for QDM at Monterey. The U-Net, by contrast, tracks the observed distribution closer in these months, matching the full shape of the observation curve.

One additional feature worth noting is the upper tail. In several months the raw forecast and observed distributions converge near the highest temperatures, suggesting the forecast captures the hottest extremes reasonably well even when it underperforms across the rest of the distribution. However, in months like January, March, April, and June, the forecast maximum does not reach the highest observed temperature at all, meaning the cold bias here is not just a shift in the center of the distribution but also a truncation of the upper end of the distribution. Neither correction method fully resolves this maximum temperature gap, pointing to a persistent limitation in capturing the most extreme coastal heat events. It is also worth noting that in some months, notably August and October, QDM shifts the distribution too far in the opposite direction, with the corrected CDF sitting to the right of the observed curve, indicating the correction has overcorrected and introduced

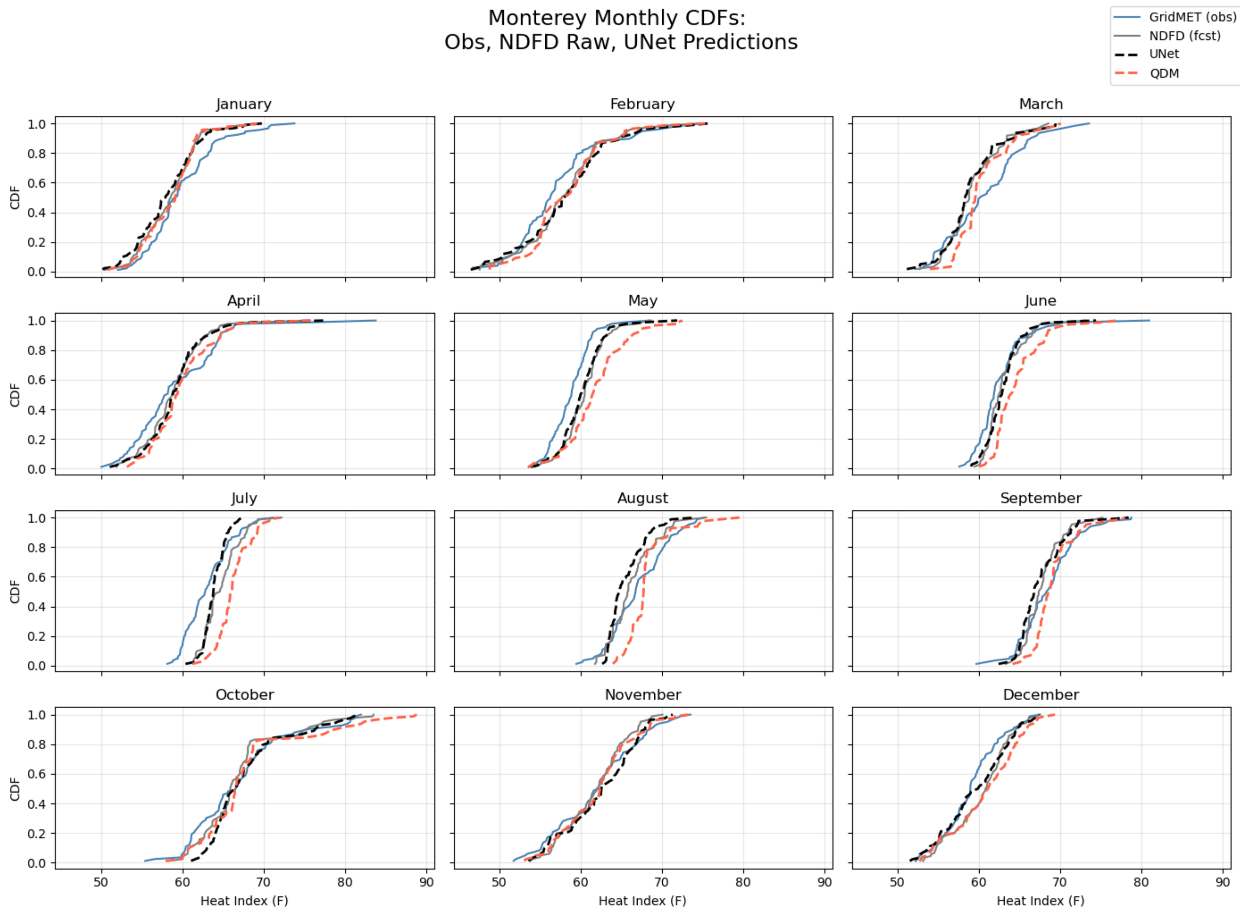

Figure 8: Monthly cumulative distribution functions (CDFs) for Monterey to compare distributional differences between approaches for the 2022–2024 testing period.

a warm bias at this location. The U-Net does not show this same overcorrection.

Extreme Error Analysis

To perform a similar extreme heat evaluation on the testing period, we first observe the full distribution of errors in Figure 9 for the testing period of 2022–2024. The pattern is generally consistent with the 2016–2018 period shown in Figure 4, with symmetric error distributions centered around zero in both raw and corrected outputs. One exception is the notably extreme tails in February, which were not present in the preliminary period. This is likely due to the fact that three years of data provides limited samples for each month that can be disproportionally influenced by anomalous days, especially in the tails. Since this pattern is not consistent across the full period, it is not interpreted as a meaningful signal.

Applying the same extreme heat evaluation to the test period, Figure 10 shows the errors on the top 5% of observed heat days across all three approaches, where error is computed as forecast minus observed heat index at each grid cell, expanding on Figure 5 for the testing period.

The summer months (June through August) follow the same general pattern shown in the pre-

Figure 9: Violin plots of full error distribution throughout the 2022–2024 test period comparing raw errors, QDM, and U-Net.

liminary analysis: raw errors are reasonably centered around zero and mostly symmetric, QDM improves this, and the U-Net tightens the distributions even further. September and October show this same pattern but on a larger scale, with more spread in the raw errors that both methods work to reduce. November and December show highly negative raw errors, QDM slightly improves these, but the U-Net shows the most improvement of the three.

The worst months are in the spring. Extreme heat days in spring produce the largest forecasting errors and are the most difficult to correct. February shows errors that are fairly uniformly distributed throughout the negative space for the raw forecast. Both QDM and the U-Net pull these toward a more normally distributed shape centered closer to zero, but heavy long tails remain in both corrected outputs, indicating that some extreme outlier days persist even after correction. March and April further highlight the difference between the two approaches. QDM makes a meaningful distributional improvement, consistent with the EMD findings, but the U-Net still outperforms it as the violin becomes noticeably smaller and more tightly centered around zero. The improvement of the U-Net over the raw errors in these months is dramatic, even if some residual error remains.

Overall, spring extreme heat days remain the hardest problem for both methods, and while neither

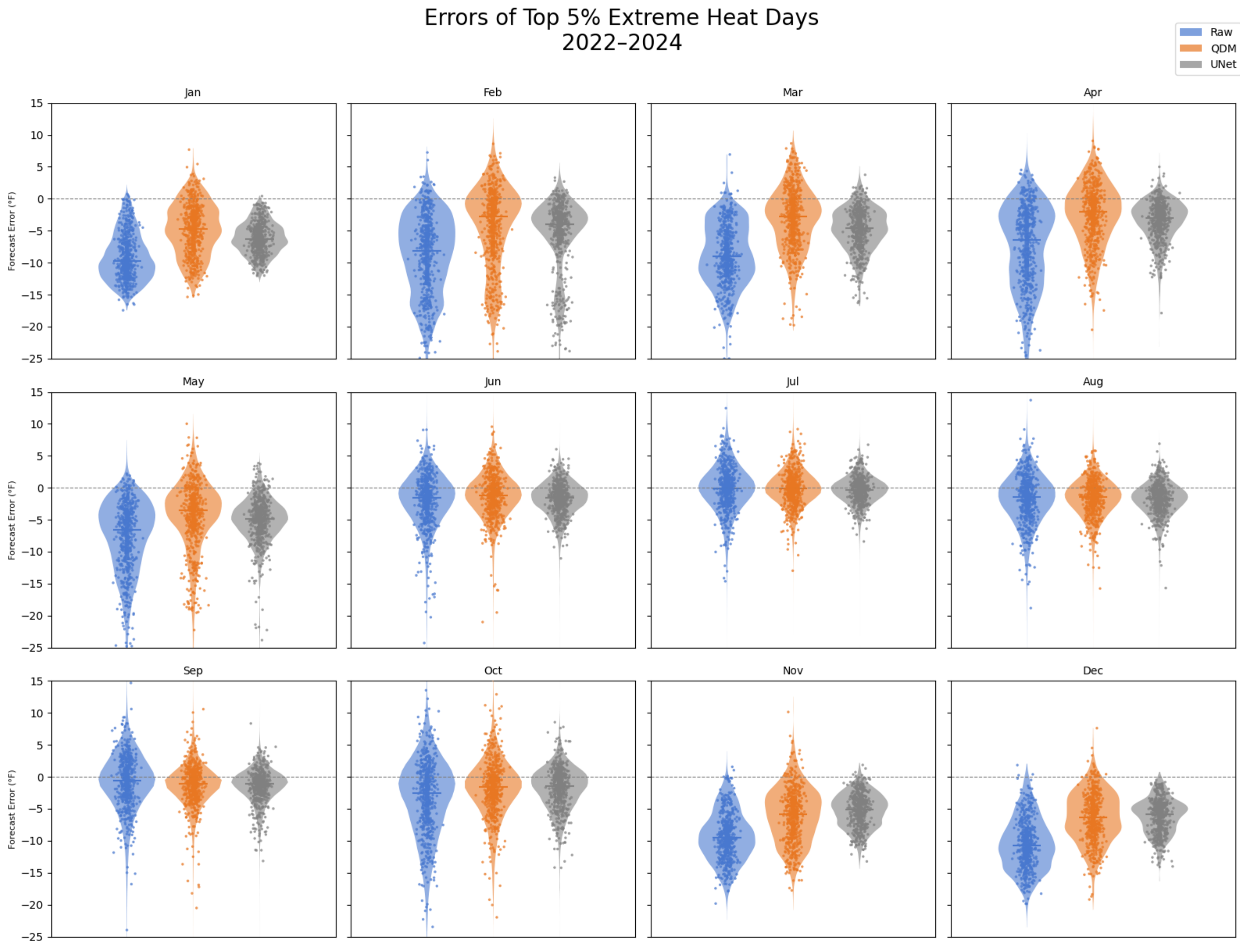

Figure 10: Violin plots of error distribution for 95th percentile of extreme heat days throughout the 2022–2024 test period comparing raw errors, QDM, and U-Net.

fully resolves the large tails in these months, the U-Net consistently comes closer.

Discussion

Consistent patterns of bias characterization and correction shortcomings have been identified in these results. The raw forecasts carry spatial and seasonal biases across California. This is observed through a persistent coastal cold bias, driven by the topographic complexity and resolution mismatches between the datasets. This is also observed through the flip to a warm bias in the Central Valley during the summer months likely due to irrigation not being represented. Both correction methods improve the raw forecast on the distributional level, but the U-Net outperforms QDM on daily accuracy throughout the test period. QDM improves the overall shape of the distribution since the EMD improves, but does not generalize when evaluated on MAE, worsening this metric by 43.0% on the test set. The U-Net reduces both metrics, suggesting it learns a better spatial and temporal correction.

The stronger performance of the U-Net compared to QDM can be attributed to differences in how the approaches work. QDM corrects each cell in isolation, using only the historical distribution of that exact location. The U-Net learns spatial corrections through convolutions, allowing it to be influenced by neighboring grid cells. This spatial awareness allows it to better understand topographically complex areas where QDM struggles. Additionally, the residual aspect of the U-Net may contribute to better generalization.

The biggest correction challenges are extreme heat days in non-summer months, which produce the largest forecast errors across both methods. Heavy, negative tails remain present even after correction in February through May. This suggests that the forecast still significantly under-predicts the heat index on these days. This directly impacts public health as early season heat events carry substantial health-risk as populations have not properly adapted to heat exposure at the start of the warm season (Lee et al. 2014).

One major limitation of both approaches is the stationarity assumption. While QDM preserves the model's projected climate change signal, it fails to capture that a model's biases may differ in the future. The corrections are learned from a fixed historical period and assume the relationship between the forecast and observed distributions remains constant over time. This assumption may not hold as climate conditions shift, which is consistent with QDM's degraded MAE performance on the test period relative to the training period. The U-Net is not immune to this problem either. Although it may generalize better than a fixed quantile mapping, the U-Net is still trained on a historical window and does not directly account for distributional shifts. Future work should explore additional bias correction approaches that can adapt to non-stationary climate conditions, such as time-varying corrections, rolling retraining on more recent data, or methods that explicitly model long-term trends in forecast bias. Incorporating additional features into the U-Net such as elevation or distance to coast may also improve correction performance where both methods currently struggle.

These results improve understanding of the systematic biases in NDFD heat index forecasts across California and how bias correction can address them. Characterizing these biases spatially and seasonally provides an understanding for where and when forecast errors exist the most. The difference in performance between QDM and the U-Net demonstrates that spatial approaches offer advantages over stationary statistical corrections. These findings have direct implications for tools like CalHeatScore, which relies on NDFD forecasts to produce daily heat-health rankings. Improving the accuracy of the underlying forecast can help improve the reliability of heat-health risk assessment and communication.

References

Abatzoglou, John T. 2013. "Development of Gridded Surface Meteorological Data for Ecological Applications and Modelling." *International Journal of Climatology* 33 (1): 121–131. https://doi.org/10.1002/joc.3413.

"CalHeatScore." 2025. California Environmental Protection Agency. Accessed December 11, 2025. https://calheatscore.calepa.ca.gov/.

Cannon, Alex J., Stephen R. Sobie, and Trevor Q. Murdock. 2015. "Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes?" *Journal of Climate* 28 (17): 6938–6959. https://journals.ametsoc.org/view/journals/clim/28/17/jcli-d-14-00754.1.xml.

Dorman, C.E., et al. 2001. "Improving Short-Term Numerical Weather Prediction in the California Coastal Zone by Dynamic Initialization of the Marine Boundary Layer." *Monthly Weather Review* 129 (2): 275–294. https://doi.org/10.1175/1520-0493(2001)129⟨0275:ISTNWP⟩2.0.CO;2.

Ghil, Michael, Myles R. Allen, Michael D. Dettinger, Kazuyuki Ide, Dmitri Kondrashov, Michael E. Mann, Andrew W. Robertson, et al. 2002. "Advanced Spectral Methods for Climatic Time Series." *Reviews of Geophysics* 40 (1): 1003. https://doi.org/10.1029/2000RG000092.

Hamill, T.M., et al. 2023. *Machine Learning for Bias Correction of NWP Forecasts.* Technical report 520. National Centers for Environmental Prediction. https://library.oarcloud.noaa.gov/noaa_documents.lib/NWS/NCEP/NCEP_office_notes/NCEP_office_note_520.pdf.

Industrial Economics, Incorporated. 2024. *Impacts of Extreme Heat to California's People, Infrastructure, and Economy: Pioneering Analysis Measuring the Uninsured and Insured Costs of Extreme Heat Events.* Sacramento, CA: California Department of Insurance, June. https://www.insurance.ca.gov/01-consumers/180-climate-change/upload/Impacts-of-extreme-heat-to-California-s-people-infrastructure-and-economy-by-California-Department-of-Insurance-June-2024.pdf.

Lee, M., F. Nordio, A. Zanobetti, P. Kinney, R. Vautard, and J. Schwartz. 2014. "Acclimatization across space and time in the effects of temperature on mortality: a time-series analysis." *Environmental Health* 13 (89). https://doi.org/10.1186/1476-069X-13-89.

Lobell, David B., Céline Bonfils, Lara M. Kueppers, and Mark A. Snyder. 2008. "Irrigation Cooling Effect on Temperature and Heat Index Extremes." *Geophysical Research Letters* 35 (9). https://doi.org/10.1029/2008GL034145. https://doi.org/10.1029/2008GL034145.

Maraun, D. 2016. "Bias Correcting Climate Change Simulations – a Critical Review." *Current Climate Change Reports* 2 (4): 211–220. https://doi.org/10.1007/s40641-016-0050-x.

National Weather Service. 2026a. "Heat Index Equation." National Oceanic and Atmospheric Administration. https://www.wpc.ncep.noaa.gov/html/heatindex_equation.shtml.

———. 2026b. "National Digital Forecast Database." Accessed March 20, 2026. https://www.weather.gov/mdl/ndfd_home.

NOAA Environmental Modeling Center. 2025. "ML4BC: Machine Learning for Bias Correction." Accessed June 4, 2026. https://github.com/NOAA-EMC/ML4BC.

Office of Environmental Health Hazard Assessment. 2026. *CalHeatScore Methods Documentation Version 2.0.* Technical report. California Environmental Protection Agency, January. https://calheatscore.calepa.ca.gov/sites/default/files/2026-02/methods_v2.pdf.

Qian, Yun, Maoyi Huang, Bo Yang, Larry K. Berg, Kai Zhang, and L. Ruby Leung. 2020. "Neglecting Irrigation Contributes to the Simulated Summertime Warm-and-Dry Bias in the Central United States." *npj Climate and Atmospheric Science* 3 (1): 1–8. https://doi.org/10.1038/s41612-020-00135-w. https://doi.org/10.1038/s41612-020-00135-w.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015,* 234–241. Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28.

Rubner, Y., C. Tomasi, and L.J. Guibas. 1998. "A metric for distributions with applications to image databases." In *Proceedings of the IEEE International Conference on Computer Vision,* 59–66. Bombay, India, January.

Teutschbein, C., and J. Seibert. 2012. "Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods." *Journal of Hydrology* 456:12–29. https://doi.org/10.1016/j.jhydrol.2012.05.052.

Teutschbein, Claudia, and Jan Seibert. 2013. "Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions?" *Hydrology and Earth System Sciences* 17:5061–5077. https://doi.org/10.5194/hess-17-5061-2013. https://doi.org/10.5194/hess-17-5061-2013.

Thrasher, Bridget, Edwin P. Maurer, Cailin McKellar, and Philip B. Duffy. 2012. "Technical Note: Bias correcting climate model simulated daily temperature extremes with quantile mapping." *Hydrology and Earth System Sciences* 16:3309–3314. https://doi.org/10.5194/hess-16-3309-2012. https://hess.copernicus.org/articles/16/3309/2012/.

Willmott, C.J., and K. Matsuura. 2005. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." *Climate Research* 30:79–82. https://doi.org/10.3354/cr030079.