

Title

When the Night Stops Cooling: Modeling Extreme Warm Nights Across California Weather Stations

Authors

Rohit Krishnan

Independent Researcher/Student

American High School

Abstract

This paper investigates which factors influence extreme warm nights across California using NOAA GHCN-Daily station observations. We defined extreme warm nights using station-specific 95th percentile summer TMIN thresholds based on a historical baseline period. For each station-year, the percentage of valid summer nights classified as extreme warm nights was calculated. A random forest regression model was then used to explore the effect of year, latitude, longitude, elevation, and distance to coast on extreme warm-night frequency. Partial dependence plots and permutation feature importance results suggest that year has a disproportionately large influence on nighttime heat across California, and that it positively influences the percent of extreme warm nights a station experiences. The next most important variable was distance to the coast, which also had a positive influence on the percent of extreme warm nights. These results showcase how both temporal changes and geographic characteristics may play an important role in explaining differences in extreme nighttime heat across California. In terms of predictive power, the random forest regressor resulted in an R^2 value of 0.410, an MAE of 2.513, and an RMSE of 3.765 when used to predict on stations within a withheld test set and considerably outperformed a naive baseline. This performance showcases how the selected geographic and temporal variables provide meaningful predictive information. Overall, these findings suggest that station geography and climate context may help identify areas where nighttime heat is more common or changing, while also providing a foundation for further investigation into the environmental factors contributing to extreme warm-night events across California.

Introduction

Within California, climate change is causing severe environmental and health risks and challenges (California Energy Commission, 2018). The state's geographical diversity, from the coast to the hot deserts in the south, means extreme heat isn't changing evenly across the state. While people focus on daytime trends, nighttime trends get less attention. Warm nights

mean humans cannot get essential recovery time, which increases health risks for vulnerable populations (U.S. Centers for Disease Control and Prevention, 2023).

Several studies indicate a strong need for real-world urban planning and public health changes to mitigate the impact of rising nighttime heat, especially for at-risk individuals (Wesley et al.). Officials need to know which geographic regions experience the highest nighttime heat exposure to prioritize cooling centers, urban growth, heat warnings, and other quality-of-life improvements, such as cooling strategies (California Natural Resources Agency, 2018).

To investigate these patterns, long-term weather station data can help give us a baseline for tracking temperature shifts. On top of that, interpretable machine learning (ML) methods can be used as tools to help scientists and researchers model complex relationships between temperature and other geographic features such as elevation. In particular, random forests (RF) (Breiman, 2001) are an attractive method for this type of analysis because they are interpretable, easy to implement and tune, and interpretation methods such as partial dependence plots (PDPs) and permutation feature importance can help provide insight into which inputs are most influential in the response.

In this study, we use an RF model to investigate how geographical features, including distance to the coast, elevation, latitude, longitude, as well as year, influence nighttime temperatures in California, USA, over the period 1980-2020. We use interpretable ML methods, including PFI and PDPs, to provide an additional level of explainability to our findings.

Methods

This study utilizes historical climate observations to analyze temperature trends across California. The primary dataset employed is the NOAA GHCN-Daily (Global Historical Climatology Network Daily) database, which provides comprehensive daily land surface observations compiled from international station networks.

To conduct this analysis, specific meteorological variables were extracted for the state of California, covering the study period from 1970 to 2025. The core variables utilized from the dataset include:

TMAX: Daily maximum temperature recorded at each station.

TMIN: Daily minimum temperature recorded at each station.

Given the seasonal focus of the research, summer is explicitly defined for this study as the four months spanning June, July, August, and September.

Raw climate datasets often contain missing values, reporting errors, or incomplete temporal coverage that can skew spatial and temporal analyses. To ensure high data integrity, a rigorous filtering process was applied to the initial pool of California stations:

1. Completeness Thresholds: We retained only stations with sufficiently complete (90%) historical records across the designated study period.
2. Seasonal Filtering: We systematically removed individual station-years with an insufficient number (less than 90%) of valid summer TMIN observations to prevent seasonal bias.
3. Anomaly Exclusion: Missing, flagged, or physically invalid temperature values were filtered out before computation.

Following the implementation of these quality control filters, a total of **230 valid stations** remained for the final analysis. For each of the 230 retained stations, core metadata parameters were compiled, including station name, latitude, longitude, and elevation. To support spatial modeling and evaluate geographic influences on temperature variations, an additional derived variable was introduced: the shortest distance to the California coast. This variable was calculated for each station relative to the coastline to account for marine moderation effects on local microclimates.



Figure 1: California station map colored by elevation or distance to coast.

The final 230 stations were distributed across diverse coastal, valley, mountain, and desert regions of California (Figure 1). Station elevations ranged from -59.1 to 3800.9 m, and distances to the coast ranged from 1.235 to 332.5 km. The spatial distribution ensures adequate representation of the state's complex topography and microclimates.

Modeling

This section outlines the mathematical definitions, quantitative metrics, and the RF model used to model extreme warm-night trends across California.

Extreme warm nights were quantified using a localized threshold approach to account for regional temperature variability. The climatological baseline period was established as 1981 to 2010. For each station, the warm-night threshold was defined as the 95th percentile of all summer minimum temperatures T_{MIN} recorded during this baseline period. Consequently, a warm night is identified whenever a summer day's T_{MIN} exceeds that specific station's 95th-percentile threshold. Station-specific thresholds were utilized because California encompasses diverse geographic and climatic zones, ranging from cool marine coastal strips and hot inland valleys to high-altitude mountain ranges and arid deserts. A fixed, statewide temperature threshold would fail to capture relative extremes accurately, as an absolute threshold that represents an extreme event in a coastal zone would be entirely ordinary in the interior desert.

To quantify the yearly occurrence of extreme heat events while accounting for variations in data availability, the annual warm-night frequency was computed for each station-year as:

$\text{percent_warm_nights} = 100 \times \frac{\text{warm_night_count}}{\text{valid_summer_TMIN_days}}$, where warm_night_count is defined as the total number of summer days within a given year where T_{MIN} exceeded the station-specific 95th percentile threshold, and $\text{valid_summer_TMIN_days}$ is the total number of summer days with non-missing, quality-controlled T_{MIN} observations.

The percentage metric was selected over raw counts because it effectively normalizes the data, preventing observational discrepancies or missing data gaps in certain station-years from skewing the longitudinal frequency trends.

To model the spatial and temporal distribution of extreme warm nights, a suite of predictor variables was compiled for each observation:

- Year: Captures temporal trends and long-term climate change trajectories over the study period.
- Latitude: Represents north-south geographic gradients and broad solar radiation patterns.
- Longitude: Captures east-west spatial variations across California's complex topography.
- Elevation (m): Accounts for lapse-rate temperature differences across varying topographic heights.
- Distance to coast (km): Measures proximity to the Pacific Ocean to quantify marine thermal moderation.

Random Forest Regression

To map the complex, non-linear relationships between climate drivers and extreme warm nights, an RF regression model was implemented. RFs work by training an ensemble of randomized decision trees and averaging their predictions, which reduces overfitting and yields more stable

predictions than a single tree (Breiman, 2001). In this study, the RF framework served as an exploratory tool to map predictive feature dependencies rather than functioning as a formal causal model. The model was trained to predict percent_warm_nights using station geography and the observation year as input features. To evaluate out-of-sample performance, the dataset was split geographically to evaluate performance on held-out stations. We used a geographic split rather than a temporal split because we are more interested in model performance over geographical features than over time. Hyperparameters, specifically the number of trees (n_estimators), maximum tree depth (max_depth), and minimum samples required to split a node (min_samples_split), were tuned using Grid Search cross-validation, grouped by station, to identify the optimal configuration that minimized prediction error. We found the best tuning parameters to be:

- Depth: trees were grown to full depth
- Maximum features used per split: all features
- Minimum samples per leaf: 2
- Number of trees:1000

The following post-hoc interpretation techniques were applied to extract physical insights:

- Permutation feature importance (PFI): This metric evaluates which variables the model relies on most heavily by randomly shuffling each predictor column and measuring the resulting drop in model accuracy. Variables that cause the largest performance degradation upon shuffling are deemed most critical to the model's predictive process.
- Partial dependence plots (PDPs): These plots illustrate the marginal relationship between a single predictor variable and the predicted warm-night frequency, holding all other features constant, thereby revealing how changes in factors like elevation or baseline temperature translate into shifts in extreme night frequency.

Model performance and predictive error were assessed using standard regression metrics:

- MAE (Mean Absolute Error): Measures the average magnitude of absolute errors.
- RMSE (Root Mean Square Error): Penalizes larger prediction errors more heavily.
- R^2 (Coefficient of Determination): Quantifies the proportion of variance in warm-night frequency explained by the feature set.

Results

This section presents the RF regression modeling and interpretability analysis using PDPs and PFI to complement the main modeling of extreme warm nights across California.

Station Geography and Coverage

The filtering and quality-control process yielded a robust final dataset containing 230 stations. To capture relative thermal extremes rather than absolute temperature cutoffs, station-specific baselines were evaluated. The station-specific warm-night thresholds varied across California,

reflecting differences between coastal, inland, mountain, and desert climates. As shown in Figure 2, the majority of station thresholds cluster between 15°C and 24°C.

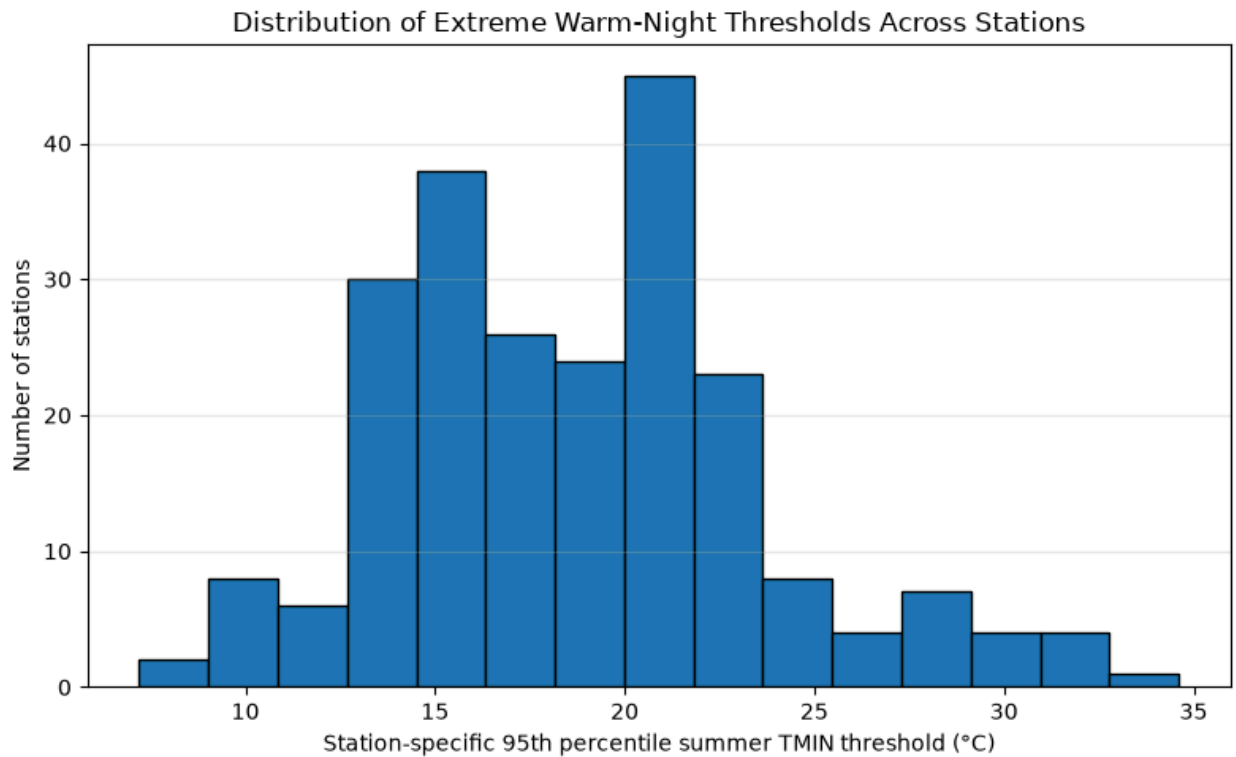


Figure 2: Frequency histogram showing station-specific 95th-percentile summer minimum temperature T_{MIN} thresholds across California stations.

Longitudinal trends in extreme heat were examined across the study period by calculating decadal trend slopes for each station. The analysis evaluated temporal trends by calculating linear regression coefficients for extreme warm-night frequency across the station network over time. As illustrated in Figure 3, the distribution of these regression slopes is heavily skewed toward positive values, with the majority of stations exhibiting a rate of change between 0 and 3 additional warm nights per decade. While minor year-to-year noise and isolated negative trend coefficients exist, the aggregate pattern demonstrates a widespread, consistent upward shift in nighttime heat extremes throughout the region.

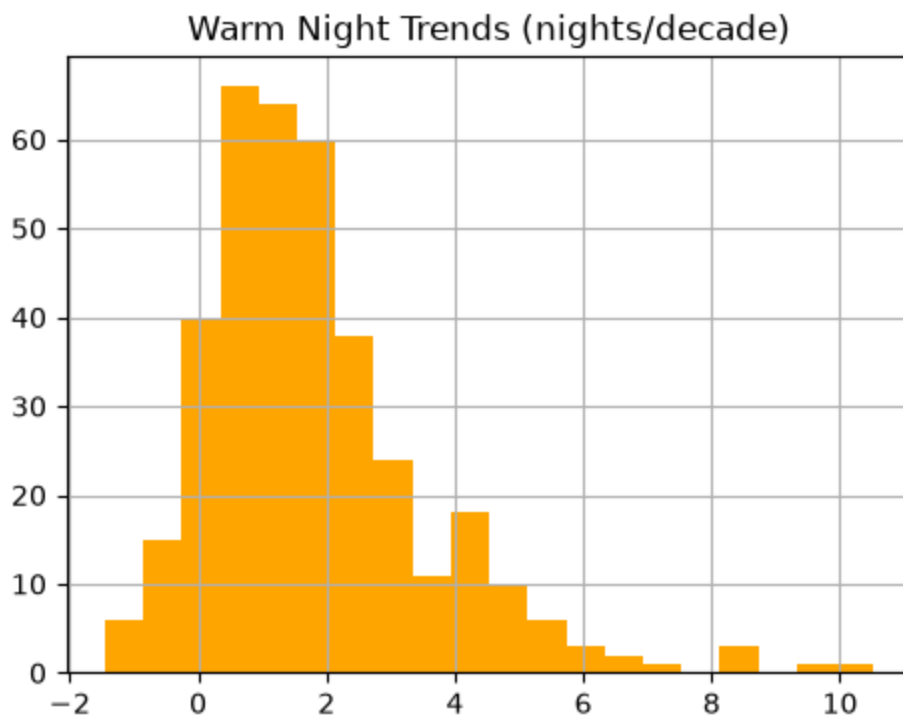


Figure 3: Distribution of station-level warm-night trends measured in nights per decade across California.

Random Forest Performance

The RF model showed moderate predictive performance ($R^2 = 0.410$), improving upon the naive baseline (the training set average). The RF model reduced RMSE from 4.913 to 3.765 relative to the baseline (Table 1). The model performed better for low and moderate warm-night frequencies but struggled with high warm-night percentages (Figure 4).

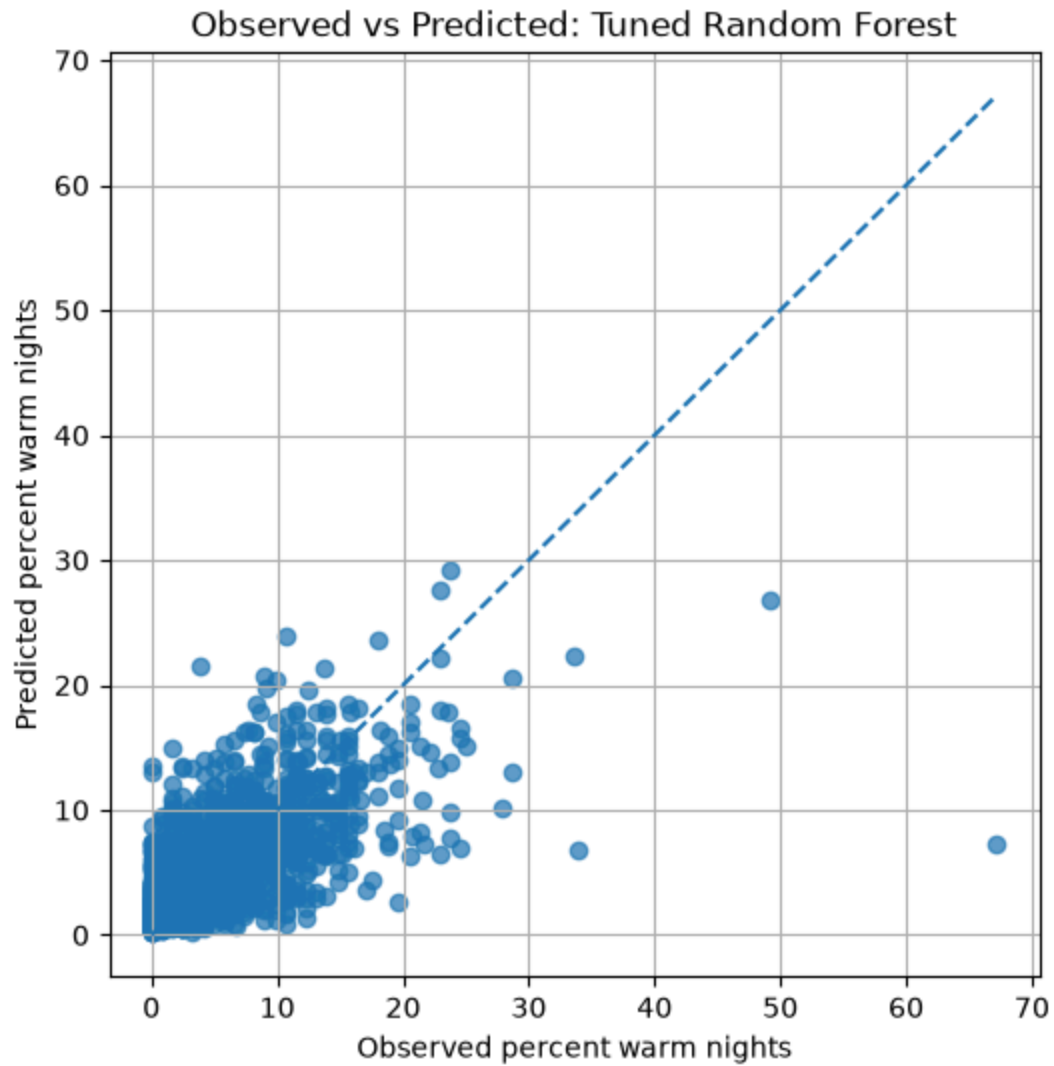


Figure 4: Observed vs predicted percent warm nights on the test set.

Table 1: Random forest regression predictive metrics on the test set compared to the baseline model (training set average of all values in percent_warm_nights).

Metric	Regression Model	Baseline Model (training set average)

MAE (%)	2.513	3.596
RMSE (%)	3.765	4.913
R²	0.410	-0.005

Model Errors

An examination of prediction residuals provides insight into systematic model biases. Residuals were more aggregated around zero for lower predicted values, but more spread out at higher predicted values. This indicates that the model generally performs well when used to predict on typical conditions, but for higher values, it struggles and underpredicts the percent of warm nights (Figure 5).

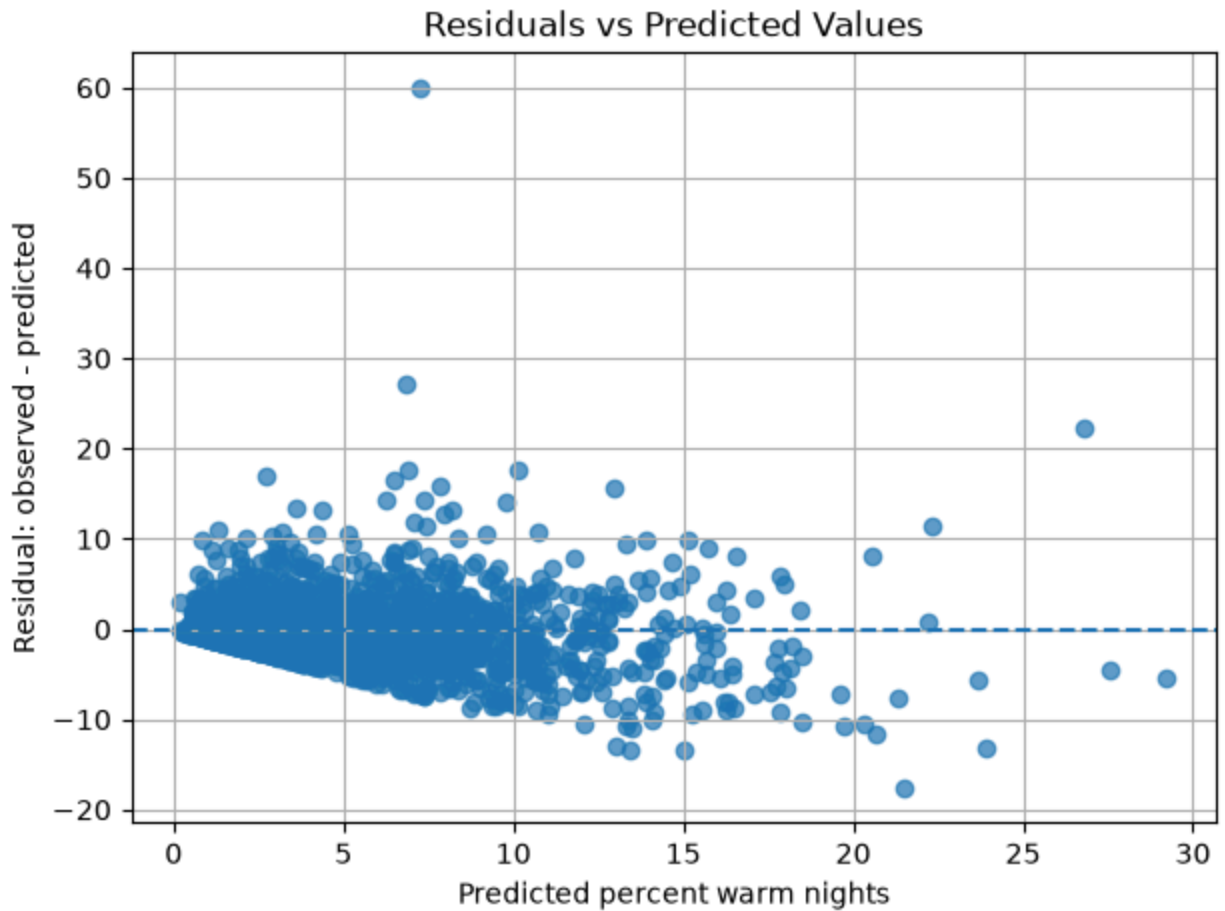


Figure 5: Residuals versus predicted percent_warm_nights.

Feature Importance and Partial Dependence

To interpret the underlying drivers learned by the ensemble model, we calculated permutation feature importance. The most important features for permutation feature importance were overwhelmingly year, followed by distance to coast, longitude, latitude, and finally elevation (Figure 6).

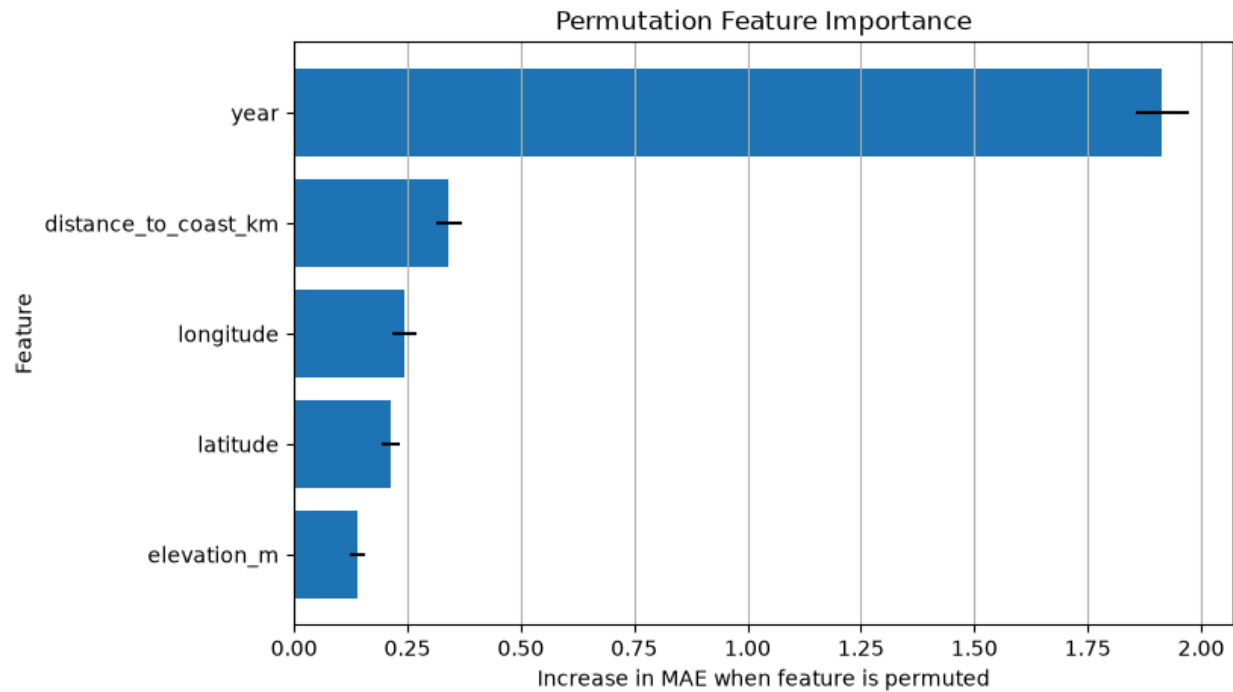


Figure 6: Permutation Feature Importance for the tuned random forest model. Values represent the mean increase in Mean Absolute Error (MAE) when each feature is randomly permuted, indicating that year is the most influential predictor, followed by distance_to_coast_km, longitude, latitude, and elevation_m.

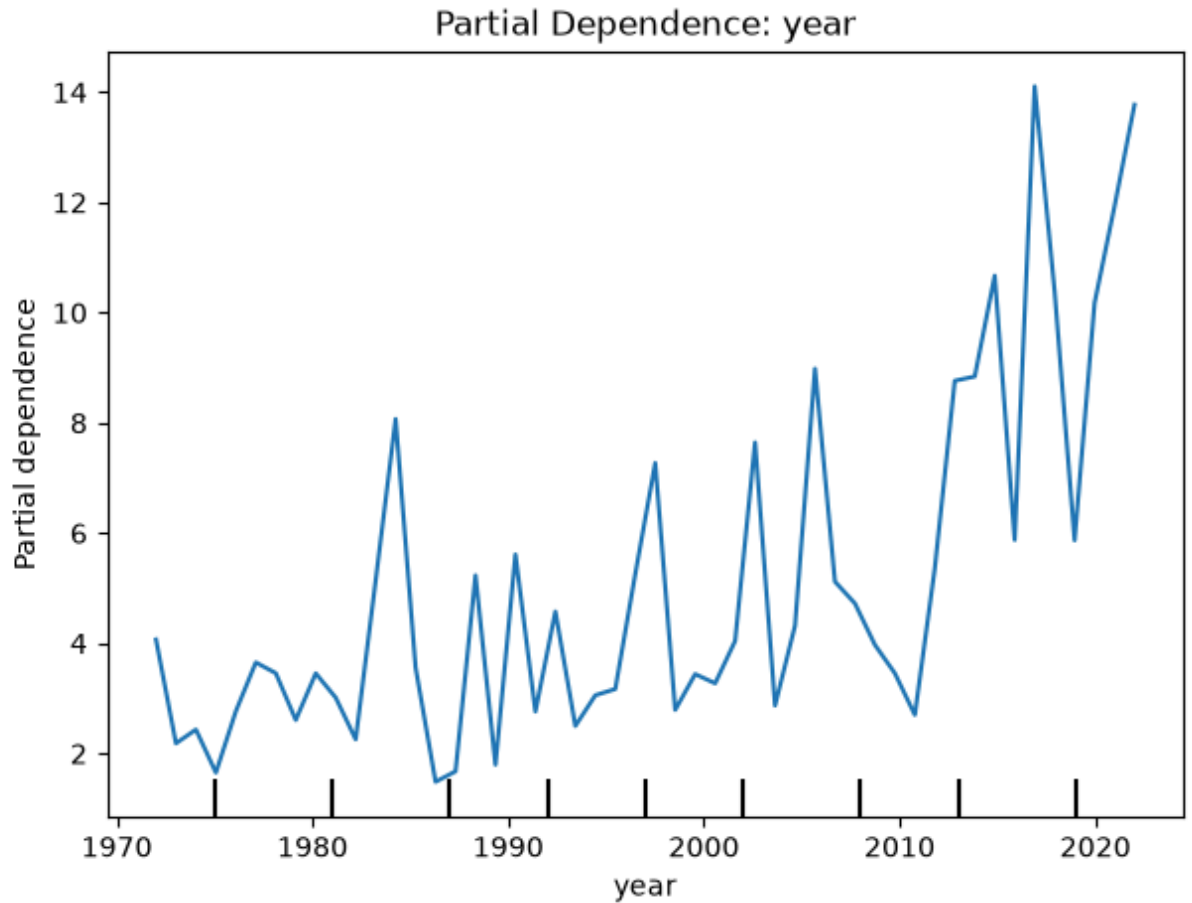


Figure 7a: Partial dependence plot for the year feature. The plot captures long-term regional warming trends across California, showing an increasing frequency of extreme warm nights in recent decades alongside sharp inter-annual climate variability (such as multi-year droughts and heat anomalies).

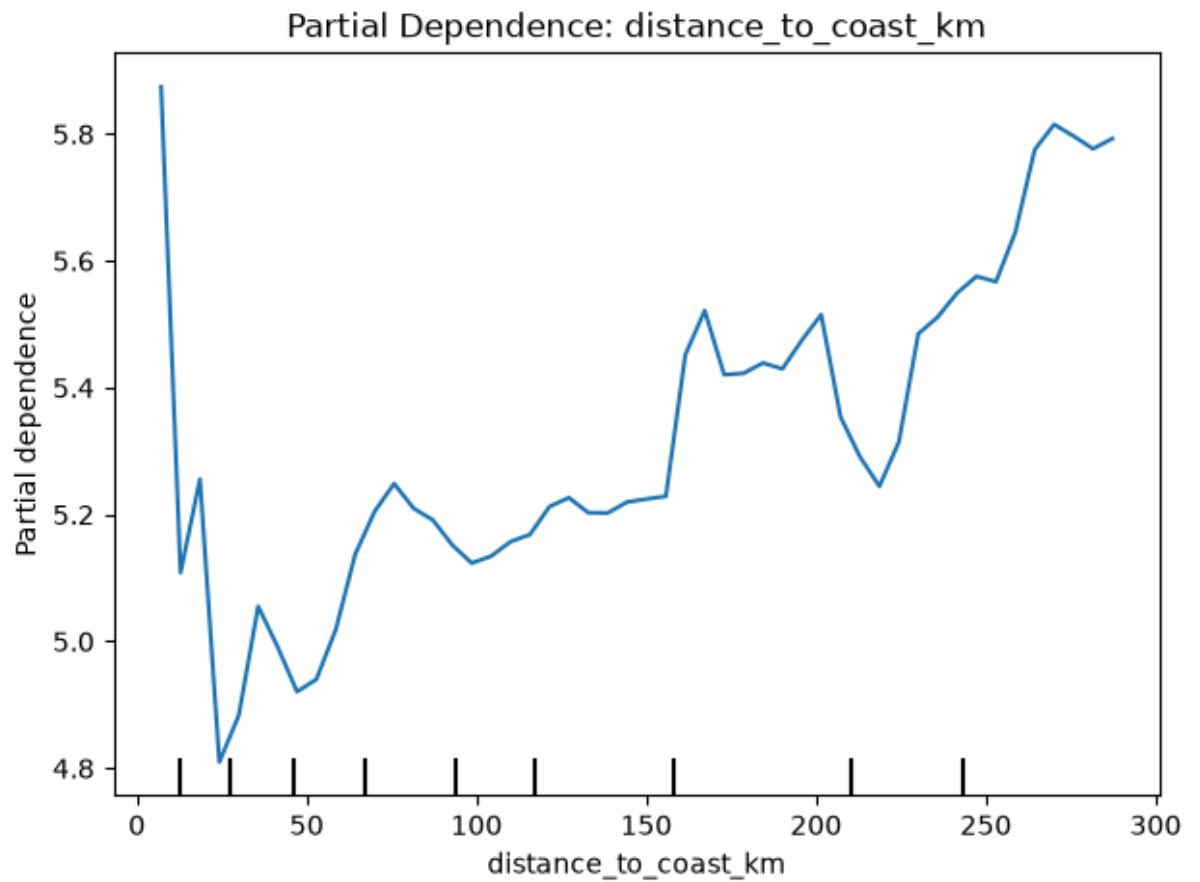


Figure 7b: Partial dependence plot for the *distance_to_coast_km* feature. The plot illustrates a sharp drop in warm-night frequency right near the coast, consistent with the cooling, temperature-buffering influence of California's marine layer, followed by an upward trend as distance increases into warmer interior valleys and continental regions.

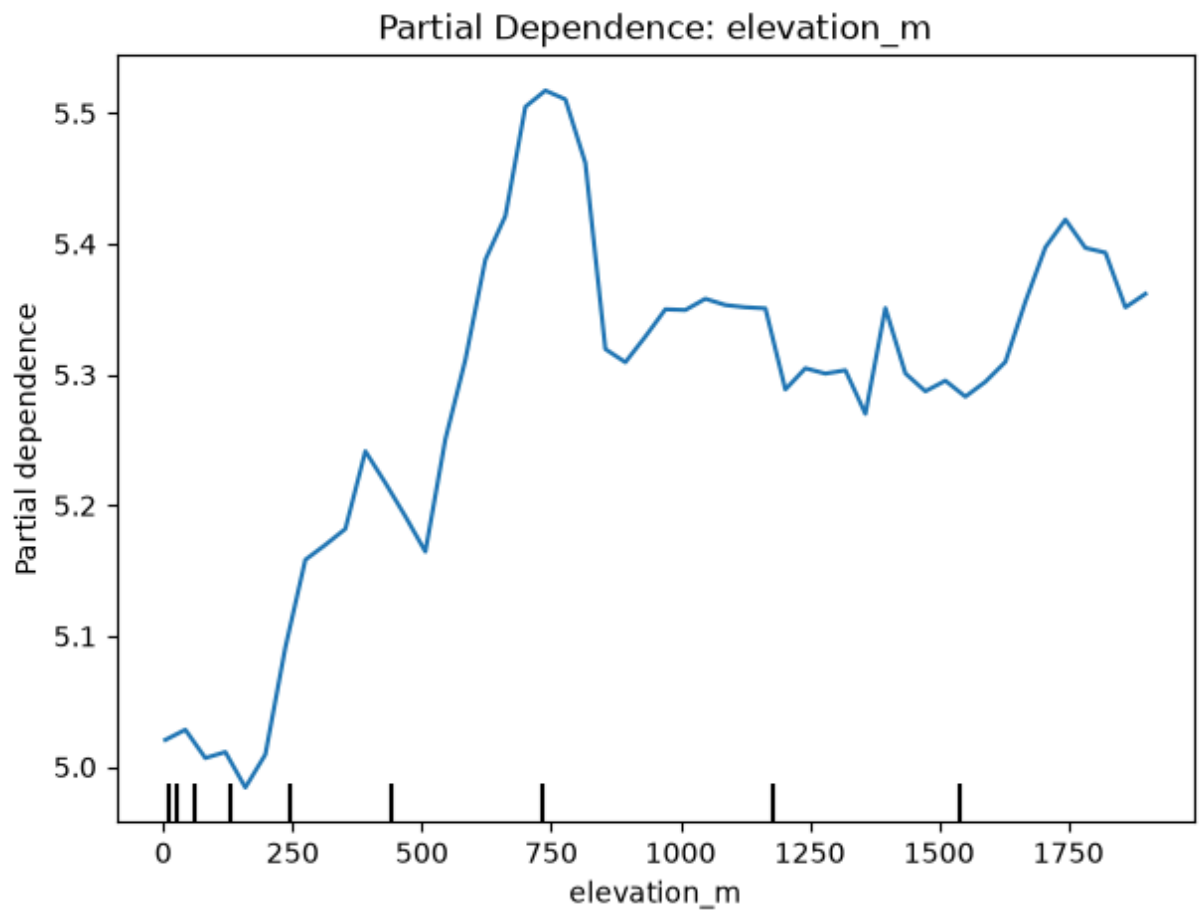


Figure 7c: Partial dependence plot for the *elevation_m* feature. The plot demonstrates an increase in warm-night frequency up to mid-elevations (around 750 meters), possibly reflecting foothill thermal belts and nocturnal temperature inversions that trap warmth, before dropping and stabilizing at higher altitudes due to naturally cooler mountain climates.

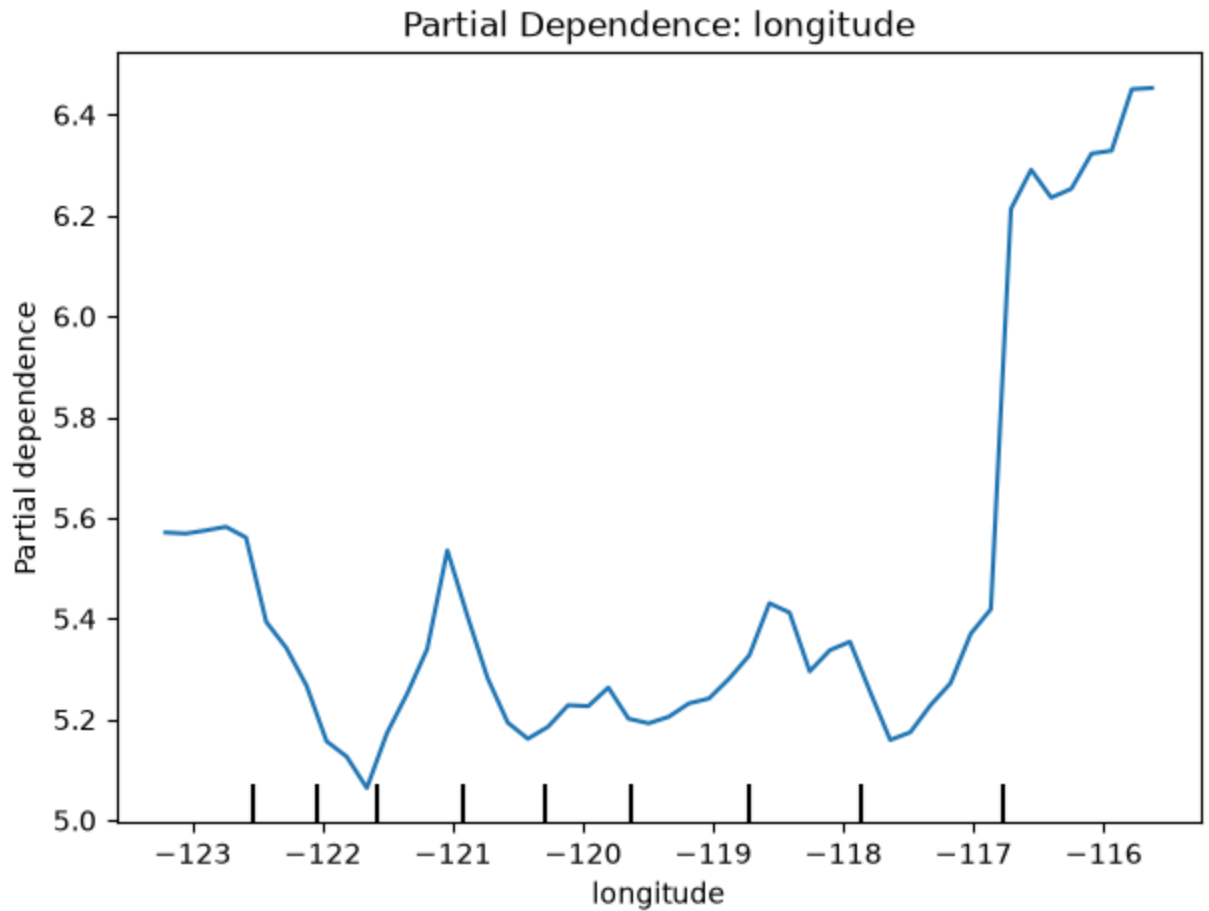


Figure 7d: Partial dependence plot for the longitude feature. The plot displays nonlinear variations across spatial coordinates, featuring a sharp upward transition as longitudes shift eastward, which could indicate the transition from coastal and valley environments into California's hot interior desert regions, where arid conditions and intense summer heating drive up the frequency of extreme warm nights.

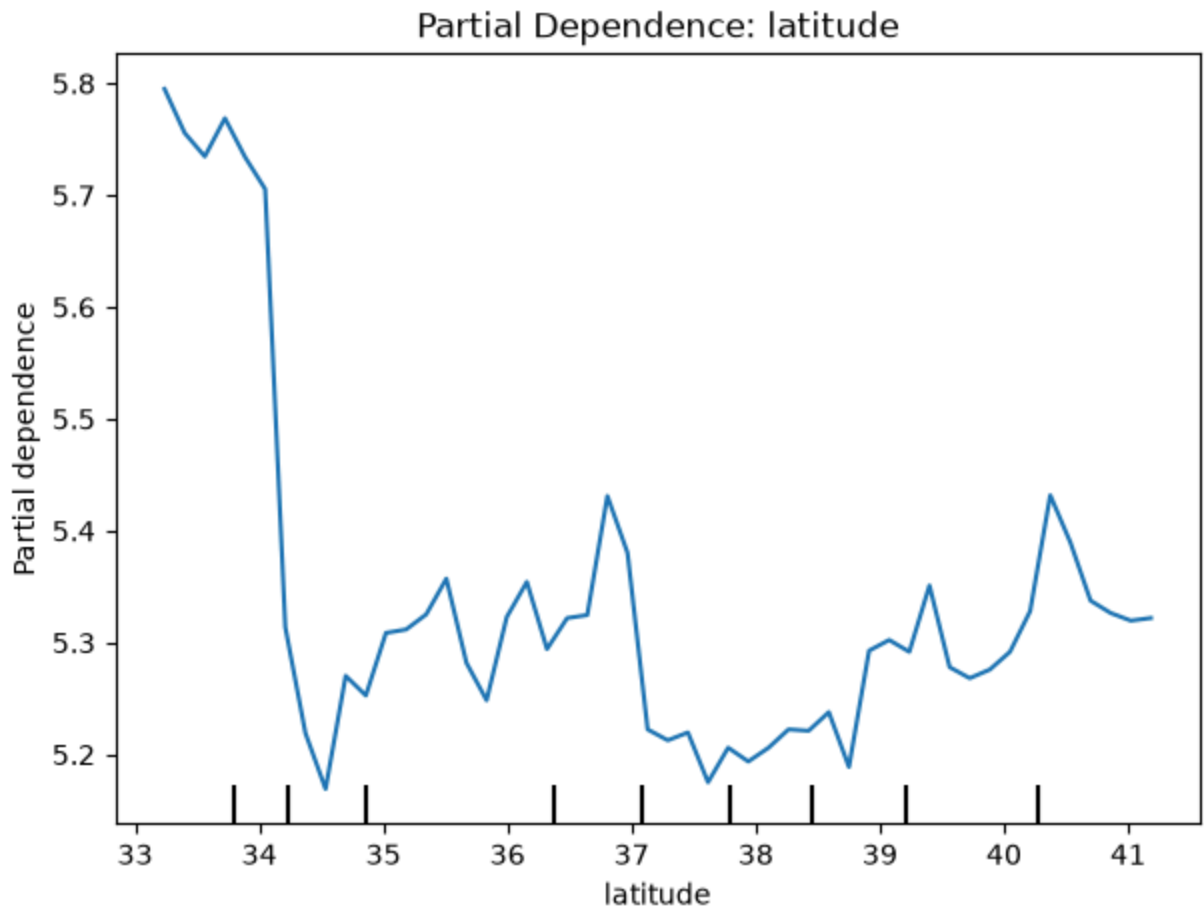


Figure 7e: Partial dependence plot for the latitude feature. While exhibiting a relatively narrow range of variation, the plot is consistent with broad geographic temperature gradients, showing that lower-latitude stations in Southern California tend to experience a slightly higher baseline percentage of warm nights before leveling off.

Discussion

This study evaluated longitudinal trends in extreme heat and assessed the predictive capacity of machine learning frameworks to model regional warm-night frequencies. The station-level trend analysis suggested that many California stations experienced increasing extreme warm-night frequency over the study period, although the magnitude of change varied by station.

The random forest model demonstrated a moderate capacity to capture general geographic patterns, outperforming the naive baseline across error metrics (reducing RMSE from 4.913 to 3.765). Furthermore, the model performed effectively for low and moderate warm-night frequencies but struggled significantly with high warm-night percentages, leading to increased error variance and underestimation of peak warm-night frequencies in anomalous years.

These modeling patterns connect directly back to California geography. Permutation importance indicated that year and distance to the coast were primary drivers, supported by partial

dependence plots showing sharp nonlinear increases as conditions transition from coastal microclimates into interior valley and desert regions. However, geography and year do not fully account for the highest extreme warm night years. This discrepancy likely stems from omitted physical or environmental controls, such as local land cover, urbanization intensity, humidity, marine layer variations, wildfire smoke, regional circulation patterns, or station-specific microclimates, many of which are correlated with year (i.e., urbanization has increased over time).

Conclusion

This type of analysis could help public health officials, city planners, or climate adaptation teams identify locations where nighttime heat is more common or increasing. Nighttime heat is critically important because people, buildings, and infrastructure have less opportunity to cool down overnight. While a station-based model cannot replace detailed local heat-risk assessments, it can help screen for broad geographic patterns and identify stations or regions that deserve closer attention.

Several methodological and data-driven limitations should be considered. The analysis relied on a limited sample of 230 stations that are unevenly distributed geographically (most are at low elevation). Additionally, while random forest regression outperformed a baseline, it struggled in predicting higher, more extreme percentages of warm nights. Finally, potentially important predictors such as land cover, population density, humidity, and urban heat island metrics were missing from the feature set, restricting the model's ability to resolve extreme localized thermal stress.

All the code, along with the paper, can be found at
<https://github.com/rohkdev/ClimateModelingResearch>

References

- **California's Fourth Climate Change Assessment.** (2018/2021). *Statewide Summary Report & Sector-Specific Reports*. California Energy Commission. Available at: <https://www.energy.ca.gov/media/2066>
- **California Legislative Analyst's Office (LAO).** *Climate Change Impacts and Public Health in California*. Report No. 4580. Available at: <https://lao.ca.gov/Publications/Report/4580>
- **California Climate Assessment Public-Health Reports.** *Cal-Heat Tool and Heat-Wave Public Health Risks*. Office of Environmental Health Hazard Assessment. Available at: https://www.energy.ca.gov/sites/default/files/2019-12/Oceans_CCCA4-EXT-2018-009_a_da.pdf
- **Breiman, L.** (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- **“Global Historical Climatology Network Daily (GHCNd).”** National Centers for Environmental Information (NCEI), Mar. 2021,

<https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily>.

- Wesley, Elizabeth Jane, et al. "Cities Are Heating Up. Better Infrastructure Can Cool Them Down." World Resources Institute, 16 July 2025, <https://www.wri.org/insights/urban-heat-effect-solutions>.