

Evaluating WRF-Derived Wet Bulb Globe Temperature Using Component Pathways, Wind-Height Adjustment, and Grid-Spacing Sensitivity

William Hatheway

hatheway.will@gmail.com

Research Article

Keywords: wet bulb globe temperature, WBGT, WRF, heat stress, Liljegren, Hunter/Minyard, Dimiceli/Piltz, FAO-56, natural wet bulb, numerical weather prediction, model validation

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11167425/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

1 **Evaluating WRF-Derived Wet Bulb Globe Temperature Using Component Pathways, Wind-Height**
2 **Adjustment, and Grid-Spacing Sensitivity**

3 William Hatheway

4 ORCID: 0000-0002-7814-5336

5 Corresponding author: William Hatheway

6 Hatheway.will@gmail.com

7 Final-review manuscript with embedded site photographs, charts, and project figures

8 Study period: 15 June-23 August 2026 | San Marcos, Texas 29.890N 97.996W

9 **Acknowledgments**

10 This research was self-funded by the author. No external funding, grants, or other financial support were
11 received for this study.

12
13 The author gratefully acknowledges Dr. Andrew J. Grundstein, Professor in the Department of Geography
14 at the University of Georgia, for his valuable guidance, advice, and scientific insight throughout the
15 development of this research. His expertise and thoughtful feedback were greatly appreciated and
16 contributed to the development and refinement of this project.

Abstract

Wet bulb globe temperature (WBGT) is widely used to evaluate environmental heat stress. However, numerical weather prediction models such as the Weather Research and Forecasting (WRF) Model do not provide WBGT as a native model variable. This study developed and evaluated a WRF-to-WBGT post-processing framework that uses standard WRF meteorological fields to calculate natural wet-bulb temperature (T_n), black-globe temperature (T_g), and dry-bulb air temperature (T_a), and then combined those components into outdoor WBGT. The framework was evaluated over 70 consecutive summer days in San Marcos, Texas (29.890N, 97.996W) from 15 June through 23 August 2026. Three component pathways were tested: Hunter/Minyard, Dimiceli/Piltz, and Liljegren. Each pathway was evaluated with four branches: Base, FAO-56 wind-height adjustment, Carter/Bernard natural-wet-bulb substitution, and Carter/Bernard plus FAO. The calculations were performed independently for 25-km, 5-km, and 1-km WRF domains and were validated against a KestrelMet 6400 WBGT station distributed through the Ambient Weather Network and a Tempest One station. Full-day performance was evaluated with bias, mean absolute error (MAE), root-mean-square error (RMSE), correlation, component diagnostics, and meteorological-driver errors; daily-maximum magnitude and timing were evaluated separately. A 10,000,000-replicate whole-day bootstrap was used to determine whether the small differences among the leading configurations were statistically resolved. Against KestrelMet/AWN observations, d01 Liljegren Carter_FAO produced the lowest full-day RMSE (3.914°F), although its advantage over d01 Hunter Carter_FAO was not statistically resolved. Against Tempest, d01 Liljegren FAO and d01 Liljegren Carter were statistically indistinguishable co-leaders. Daily maxima favored different domain-method combinations, and the best peak RMSE differences were also generally unresolved. WRF showed persistent dry, high-wind, and positive shortwave-radiation errors; these errors affected T_n and T_g differently and sometimes shown in the final WBGT. The results support Liljegren as the preferred general full-day framework, while Hunter Carter_FAO remains a strong alternative, particularly for daily maxima. The study also shows that component diagnostics, observation source, and local exposure must be considered before declaring one WRF-derived WBGT configuration superior.

Keywords: wet bulb globe temperature; WBGT; WRF; heat stress; Liljegren; Hunter/Minyard; Dimiceli/Piltz; FAO-56; natural wet bulb; numerical weather prediction; model validation

1. Introduction

Environmental heat stress depends on more than air temperature. Humidity controls the ability of the body to cool through evaporation, wind affects convective and evaporative heat transfer, and solar and longwave radiation can add substantial thermal loading during outdoor activity. Wet bulb globe temperature (WBGT) combines these effects into a practical heat-stress index and is used in athletics,

occupational safety, military operations, and other outdoor activities where heat exposure can affect health and performance (Budd 2008; Grundstein et al. 2015).

Direct WBGT measurements are useful, but continuous observations are not available everywhere that heat-risk guidance may be needed. For that reason, several studies have developed methods for estimating WBGT from standard meteorological measurements. Hunter and Minyard (1999) developed an operational approach that estimates WBGT from routine weather observations, Liljegren et al. (2008) developed a more physically based method that explicitly represents heat and mass transfer, and Dimiceli, Piltz, and Amburn (2011) proposed a simplified black-globe-temperature calculation that can be paired with other wet-bulb treatments. These approaches make it possible to estimate WBGT from weather-station observations and from numerical weather model output.

Using a gridded model output introduces another level of uncertainty. The Weather Research and Forecast Advanced Research WRF (WRF-ARW) can simulate the temperature, moisture, wind, pressure, radiation, cloud, and land-surface fields needed for a WBGT calculation, but WRF does not directly simulate the point exposure measured by a WBGT instrument. A model grid cell may represent conditions over several square kilometers, while an instrument, by comparison, responds to the immediate effects of grass, trees, fences, nearby structures, and local airflow. This difference matters because WBGT is sensitive to wind and radiation as well as temperature and humidity. Therefore, a WRF-derived WBGT value can be physically reasonable while individual meteorological drivers still contain meaningful errors.

Previous WBGT research supports this concern. Carter et al. (2020) compared several methods for estimating WBGT from remote and lower-cost observations and showed that the treatment of wind, radiation, and natural wet-bulb temperature can substantially change the result. Clark and Konrad (2020, 2024) also demonstrated that WBGT varies across observing environments and that point measurements can differ from gridded or nearby estimates because of local exposure. These studies make two issues especially important for numerical-model validation: the WBGT calculation method must be identified clearly, and the observation source must be treated as part of the uncertainty rather than assumed to be a perfect reference.

In this study, those issues are addressed with a controlled WRF-to-WBGT experiment in central Texas. One WRF physics configuration was held fixed for the entire study period; the post-processing pathway, correction branch, model domain, and observation source were then evaluated systematically. Three primary component pathways were used: Hunter/Minyard, Dimiceli/Piltz, and Liljegren. Four branches were applied within each pathway: Base, FAO-56 wind-height adjustment, Carter/Bernard natural-wet-bulb substitution, and Carter/Bernard plus FAO. The same calculations were repeated for d01, d02, and d03 so that 25-km, 5-km, and 1-km WRF output could be compared without assuming that the finest grid must perform best.

The main questions were straightforward: (1) Which WRF-derived WBGT pathways provide the best agreement with field observations? (2) Does converting WRF 10-m wind to an approximate 2-m exposure wind with FAO-56 improve the WBGT calculation? (3) Does the Carter/Bernard Tn substitution improve agreement relative to each pathway baseline? (4) Does the preferred pathway or branch change with WRF grid spacing? (5) Which meteorological drivers explain the largest WBGT errors? (6) Are differences among model configurations larger than the uncertainty introduced by using different observation sources? These questions were evaluated with continuous full-day statistics, daily-maximum magnitude and timing, component diagnostics, meteorological-driver errors, and whole-day bootstrap resampling.

2. Data and Methods

2.1 Study design and validation period

The study evaluated 70 consecutive validation days from 15 June through 23 August 2026. For each date, WRF was initialized from the 00 UTC Global Forecast System (GFS) cycle and run through forecast hour 24. The production workflow downloaded GFS f000-f024 data, ran WPS and WRF, extracted the validation point from each WRF domain, calculated all WBGT pathways and branches, retrieved station observations, and merged the model and observations on a common 15-min grid. The validation window was 0700-1900 Central Time, inclusive, which produced 49 target timestamps per day. With three WRF domains, the completed project contained 10,290 model-domain validation rows.

The experimental design was intentionally kept fixed. WRF physics, domain geometry, initialization source, and the post-processing code did not change from day to day, only the atmospheric conditions changed. Twelve WRF-derived WBGT configurations were calculated in each domain, giving 36 model candidates for each observation source. This design allowed differences in performance to be traced to the component pathway, the correction branch, the WRF grid spacing, or the observation source rather than to a changing model configuration.

2.2 Study site

The validation site was located in San Marcos, Texas, near 29.8899°N, 97.9961°W. The instruments were positioned over maintained grass in a sheltered residential exposure. Trees, fencing, and a nearby house partially restrict airflow and can affect the shortwave and longwave radiation environment. The site therefore represents a practical outdoor exposure rather than an ideal open-field reference site.

This siting difference is important for interpreting WRF point validation. A WRF grid cell cannot resolve an individual fence, a tree canopy edge, or house-scale shading; the station responds to all three. The KestrelMet manufacturer also notes that nearby trees and structures can affect wind and radiation measurements and recommends open exposure when possible. This study therefore treats site sheltering as a known representativeness limitation in this paper. A more detailed microclimate analysis is reserved

118 for follow-on work. Figure 1 documents the residential exposure, the instrument placement, and the
119 surrounding backyard setting used for validation.

Validation site and instrument exposure, San Marcos, Texas

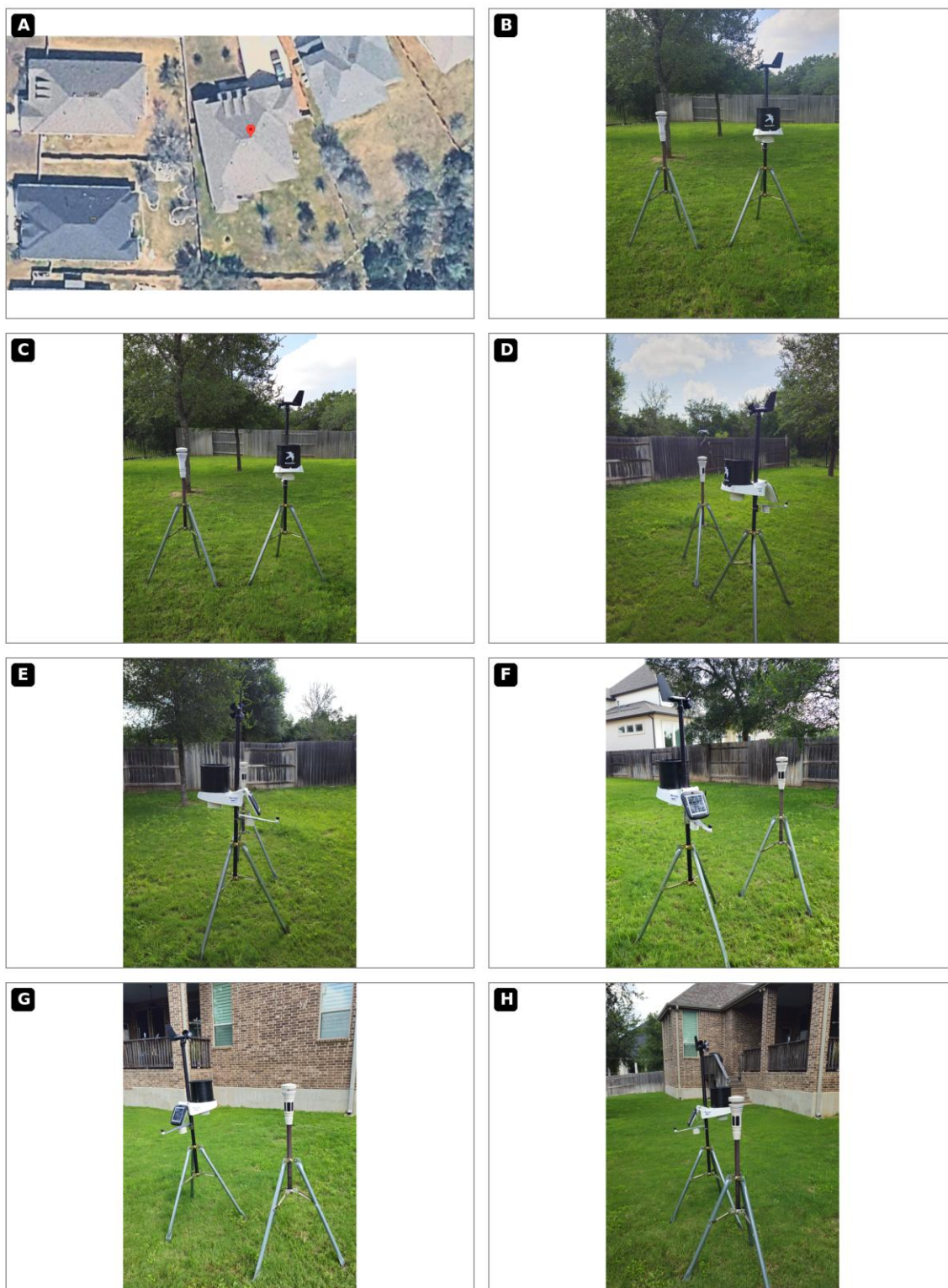


Fig. 1 Validation site and instrument exposure in San Marcos, Texas. (A) Aerial view of the residential property and surrounding neighborhood. (B-H) Ground-level views of the KestrelMet/AWN and Tempest stations installed over maintained grass within the backyard. The photographs illustrate nearby fencing, mature trees, and the adjacent residence, all of which contribute to the partially sheltered exposure discussed in Section 2.2

2.3 Observational datasets

2.3.1 KestrelMet 6400 WBGT / Ambient Weather Network

The primary observation source was a KestrelMet 6400 WBGT Wi-Fi Weather Station distributed through the Ambient Weather Network (AWN). The station is designed as a dedicated WBGT-capable weather system and provides WBGT, natural wet-bulb temperature, black-globe temperature, air temperature, relative humidity, pressure, wind, and solar radiation when the pyranometer data are available. These measurements were used as the principal reference because the station provides the WBGT components directly rather than reconstructing them only from standard weather variables (Nielsen-Kellerman 2025).

Historical AWN observations were retrieved through the Ambient Weather API. A 10-min buffer was requested around the 0700-1900 validation period so that records near each 15-min target time could be matched reliably. The observations were aligned to the target grid with nearest-neighbor matching and a tolerance of 7 min 30 s. Source timestamps were retained in the merged files so that match timing could be checked during quality control.

2.3.2 Tempest One

A Tempest One station was used as an independent secondary observation source (WeatherFlow-Tempest 2026). Tempest directly measures air temperature, relative humidity, station pressure, wind, and solar radiation. It also provides derived quantities, including dew point, wet-bulb temperature, and WBGT. Because those WBGT-related fields are derived rather than measured with a dedicated black-globe and natural-wet-bulb assembly, Tempest was not treated as interchangeable with the KestrelMet. Instead, it was used to determine how observation choice affects model validation and method ranking.

Tempest observations were requested at 1-min resolution for 0700-1900 local time. The fields used in this study were air temperature, relative humidity, dew point, wet-bulb temperature, WBGT, mean wind speed, station pressure, sea-level pressure where needed for reference, and solar radiation. The same 15-min target grid and 7.5-min nearest-match tolerance were used so that the two observation sources could be compared on a consistent time base.

2.3.3 Observation completeness and source separation

The target grid contained 3,430 timestamps for each observation source (70 days x 49 times per day). Missing station data was not filled in. Ambient WBGT, natural wet bulb, and black globe were complete at all 3,430 target times; Ambient solar radiation was available at 3,419 times. Tempest WBGT and the principal derived fields were available at 3,393 target times. Missing observations were excluded only from the statistics that required the missing variable; model values were not discarded simply because another observation field was unavailable.

2.4 WRF configuration

The atmospheric simulations used WRF-ARW version 4.8.0 and WPS version 4.7.0 (Skamarock et al. 2019). Three one-way nested domains were configured with horizontal grid spacings of 25 km (d01), 5 km (d02), and 1 km (d03). Each daily run was initialized at 00 UTC with NOAA GFS 0.25° forcing and extended through forecast hour 24; WRF history files were written every 15 min. The analysis window from 0700 to 1900 Central Time therefore sampled approximately forecast hours 12-24, depending on daylight-saving conversion and the exact UTC/local-time relationship.

The parent-domain time step was 150 s, with 5:1 parent-to-child time-step ratios. GFS forcing was supplied hourly. The physics configuration used Thompson microphysics (Thompson et al. 2008), RRTMG longwave and shortwave radiation (Iacono et al. 2008), the MYNN surface-layer and planetary-boundary-layer schemes (Nakanishi and Niino 2009), Noah-MP land-surface physics (Niu et al. 2011), and multi-scale Kain-Fritsch cumulus parameterization on all three domains (Kain 2004); urban physics was disabled. The configuration was held constant over the 70-day study so that this paper evaluates the WRF-to-WBGT framework rather than a collection of changing WRF physics experiments.

The three grid spacings were included as a sensitivity test. The study did not assume beforehand that d03 would provide the best point validation. Finer WRF grids can represent smaller-scale gradients and convective structure more explicitly. However, they can also produce sharper timing errors in clouds, radiation, and local wind. A 1-km grid also remains much larger than the immediate environment around the instruments. For this reason, d01, d02, and d03 were evaluated independently.

Table 1. Principal WRF/WPS configuration used for the 70-day experiment.

Component	Configuration
WRF / WPS versions	WRF-ARW 4.8.0 / WPS 4.7.0
Initialization	Daily 00 UTC GFS 0.25°; f000-f024; hourly forcing
Forecast length	24 h; validation uses f012-f024
Domains	d01 = 25 km; d02 = 5 km; d03 = 1 km; one-way nesting
Grid dimensions	100 x 100; 101 x 101; 101 x 101

Component	Configuration
Vertical grid	50 levels; model top 30 hPa
Projection	Lambert conformal; ref_lat 29.890; ref_lon/stand_lon -97.996
Time step	150 s parent; parent_time_step_ratio 1, 5, 5
History output	15 min, all domains; one valid time per output file
Microphysics	Thompson (mp_physics = 8)
Radiation	RRTMG longwave and shortwave (4/4); radt = 5 min
Surface layer / PBL	MYNN / MYNN (5/5)
Land surface	Noah-MP (4); 4 soil layers
Cumulus	Multi-scale Kain-Fritsch (11) on d01/d02/d03
Urban physics	Off
Nest feedback	Off (feedback = 0)
Nudging / DA	None in Paper 1

2.5 WRF-to-WBGT post-processing

For each domain, the post-processing workflow extracted the WRF grid point nearest the validation coordinates. The required fields were 2-m air temperature (T2), a 2-m dew-point diagnostic from wrf-python (td2), surface pressure (PSFC), 10-m wind components (U10 and V10), cosine of the solar zenith angle (COSZEN), and downward shortwave radiation (SWDOWN). Wind speed was calculated from U10 and V10. These variables supplied the meteorological forcing used by the component calculations.

Additional WRF diagnostics were exported when available, including Q2, GLW or LWDNB, SWDNB, SWUPB, ALBEDO, TSK, HFX, LH, QFX, PBLH, UST, VEGFRA, LAI, and LU_INDEX. These fields did not change the WBGT calculations, they were retained so that model-driver and land-surface behavior could be examined when interpreting errors. A schematic overview of the analysis chain is shown in Figure 2.

WRF-derived WBGT processing workflow used in Paper 1

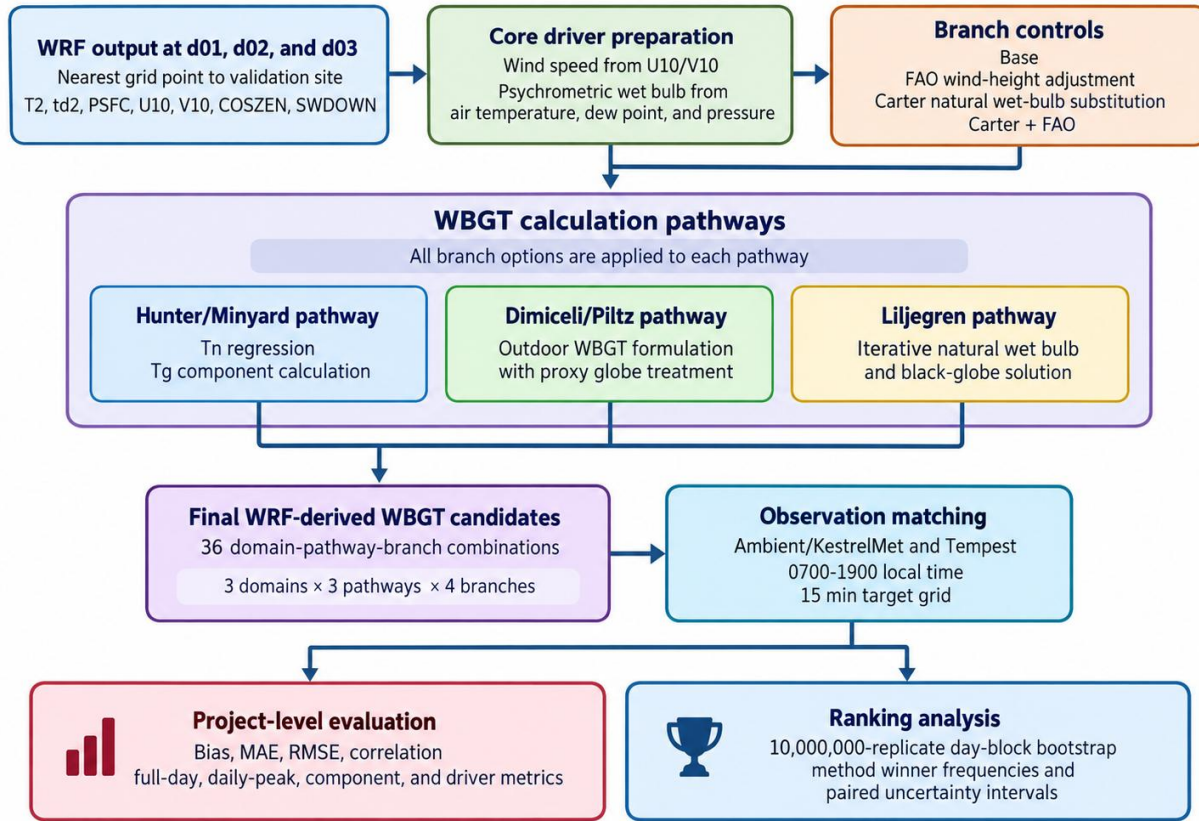


Fig. 2 WRF-derived WBGT processing workflow used in Paper 1. WRF meteorological fields from d01, d02, and d03 were sampled at the grid point nearest the validation site, converted to core driver variables, routed through psychrometric wet-bulb and wind treatments, passed through the Hunter/Minyard, Dimiceli/Piltz, and Liljegren component pathways, combined with Base, FAO, Carter, and Carter_FAO branches, and then compared with observations using project-level statistics and bootstrap ranking

2.6 Psychrometric wet bulb and outdoor WBGT closure

Psychrometric wet-bulb temperature (Tpwb) was calculated from surface pressure, air temperature, and dew point with MetPy. Tpwb was used as the input wet-bulb quantity in the Hunter/Minyard Tn regression, as the Base/FAO wet-bulb proxy in the Dimiceli/Piltz pathway, and as an input to the Carter/Bernard Tn substitution. Keeping this quantity separate from natural wet-bulb temperature is important because the two temperatures represent different exposure conditions.

Outdoor WBGT was calculated the same way for every pathway and branch:

$$\text{WBGT} = 0.70 \text{ Tn} + 0.20 \text{ Tg} + 0.10 \text{ Ta} \quad (1)$$

All component temperatures were calculated internally in degrees Celsius and converted to degrees Fahrenheit for the primary heat-stress comparisons. The 0.70 weighting on T_n makes the natural wet-bulb component the dominant term in outdoor WBGT. As a result, a change in the T_n calculation can have a much larger effect on final WBGT than a similar change in T_a or T_g.

2.7 Component pathways

2.7.1 Hunter/Minyard pathway

The Hunter/Minyard pathway was used as the empirical, operationally oriented approach. Natural wet-bulb temperature followed the Hunter/Minyard regression, which uses psychrometric wet bulb, downward shortwave radiation, and wind speed. The original Hunter/Minyard work was developed so that WBGT could be estimated from routine meteorological observations rather than relying on repeated manual field measurements:

$$T_n = T_{pwb} + 0.021 S - 0.42 u + 1.93 \quad (2)$$

where T_{pwb} and T_n are in degrees Celsius, S is shortwave radiation in W m⁻², and u is wind speed in m s⁻¹. The Hunter/Kuehn globe temperature was calculated with an iterative radiative-convective energy balance. The Hunter globe calculation used a fixed direct/diffuse shortwave partition of 0.67/0.33, atmospheric emissivity based on vapor pressure, and wind-dependent convection. This implementation preserves the published operational logic while allowing the component errors to be examined separately.

2.7.2 Dimiceli/Piltz pathway

Dimiceli/Piltz was treated as a black-globe-temperature pathway rather than a complete independent WBGT method. Its implementation used the published linearized T_g approximation, with atmospheric emissivity, solar geometry, shortwave radiation, and wind-dependent convection. A complete WBGT estimate still requires a T_n value. Therefore, the Base and FAO branches paired Dimiceli/Piltz T_g with psychrometric wet bulb, while the Carter and Carter_FAO branches paired the same T_g pathway with Carter/Bernard T_n. This distinction is maintained throughout the paper because it directly affects how the results should be interpreted.

2.7.3 Liljegren pathway

The Liljegren pathway represented the physically based WBGT formulation described by Liljegren et al. (2008). Separate iterative heat- and mass-transfer calculations were used for the wetted wick and the black globe. The calculation accounts for wind-dependent convection, atmospheric emissivity, solar geometry, radiation exchange, evaporation, temperature-dependent air properties, and sensor geometry. This makes Liljegren the most physically complete of the three pathways evaluated here.

Unlike the fixed shortwave partition used in the Hunter-based globe calculation, Liljegren calculates a dynamic direct fraction from the modeled shortwave radiation, solar zenith angle, and Earth-Sun distance. The direct and diffuse portions of the solar field therefore change with time rather than remaining fixed. This difference is especially relevant for WBGT because both the wetted wick and black globe respond to the radiative environment.

2.8 Correction and sensitivity branches

2.8.1 FAO-56 wind-height adjustment

The FAO branch tested a simple but important question: does reducing WRF 10-m wind toward an approximate 2-m exposure wind improve the WBGT calculation? We applied the FAO-56 short-grass wind-height relationship (Allen et al. 1998; Carter et al. 2020):

$$U_2 = U_z [4.87 / \ln(67.8 z - 5.42)] \quad (3)$$

where U_z is wind speed at height z (m), and U_2 is the estimated 2-m wind. For WRF, $z = 10$ m. The adjusted wind was used consistently wherever wind entered the T_n or T_g calculation for that branch. This is a height conversion over short grass, it is not a site-specific correction for trees, fences, buildings, or other local sheltering.

2.8.2 Carter/Bernard natural-wet-bulb substitution

Carter/Bernard was implemented as a controlled natural-wet-bulb substitution within each pathway following Carter et al. (2020). It does not represent a fourth WBGT family. The Carter branch changes T_n while retaining the pathway-specific T_g ; Carter_FAO makes the same T_n substitution after the FAO wind adjustment. This design allows the influence of T_n to be tested without changing the black-globe method at the same time:

$$T_n = T_a - C(T_a - T_{pwb}) \quad (4)$$

with $C = 0.85$ for $u < 0.03$ m s⁻¹, $C = 1.0$ for $u > 3$ m s⁻¹, and $C = 0.96 + 0.069 \log_{10}(u)$ otherwise. For $T_g - T_a \geq 4^\circ\text{C}$,

$$T_n = T_{pwb} + 0.25(T_g - T_a) + E \quad (5)$$

where $E = 1.1$ for $u < 0.1$ m s⁻¹, $E = 0.10/u^{1.1} - 0.2$ for $0.1 \leq u \leq 1$ m s⁻¹, and $E = -0.1$ for $u > 1$ m s⁻¹.

The Carter branch used WRF 10-m wind; Carter_FAO used FAO-adjusted wind. Because T_g is unchanged by Carter substitution, Base and Carter share the same T_g within a pathway, and FAO and Carter_FAO share the same T_g . That expected behavior was also used as an internal consistency check.

268 *Table 2. WBGT component pathways and branch definitions.*

Pathway / branch	Tn treatment	Tg treatment	Wind treatment
Hunter Base	Hunter/Minyard empirical Tn	Hunter/Kuehn iterative globe balance	WRF 10 m
Hunter FAO	Hunter/Minyard empirical Tn	Hunter/Kuehn iterative globe balance	FAO 2 m
Hunter Carter	Carter/Bernard Tn	Hunter/Kuehn iterative globe balance	WRF 10 m
Hunter Carter_FAO	Carter/Bernard Tn	Hunter/Kuehn iterative globe balance	FAO 2 m
Dimiceli/Piltz Base	Psychrometric wet-bulb proxy	Dimiceli/Piltz linearized Tg	WRF 10 m
Dimiceli/Piltz FAO	Psychrometric wet-bulb proxy	Dimiceli/Piltz linearized Tg	FAO 2 m
Dimiceli/Piltz Carter	Carter/Bernard Tn	Dimiceli/Piltz linearized Tg	WRF 10 m
Dimiceli/Piltz Carter_FAO	Carter/Bernard Tn	Dimiceli/Piltz linearized Tg	FAO 2 m
Liljegren Base	Iterative heat/mass-transfer Tn	Iterative heat-transfer Tg	WRF 10 m
Liljegren FAO	Iterative heat/mass-transfer Tn	Iterative heat-transfer Tg	FAO 2 m
Liljegren Carter	Carter/Bernard Tn	Liljegren Tg	WRF 10 m
Liljegren Carter_FAO	Carter/Bernard Tn	Liljegren Tg	FAO 2 m

269

270 **2.9 Matching, quality control, and missing data**

271 For each validation date, the study merged WRF output from d01, d02, and d03 with the two

272 observation sources on the common 15-min local-time grid. Both UTC and local timestamps were

273 retained. Project-level quality control checked for duplicate validation-date/time/domain combinations,

274 missing WRF fields, incomplete domain coverage, negative forecast lead times, and mismatches between

275 the local valid date and the assigned validation date. All 70 daily files passed the final project-level

276 structural checks.

Missing observations were handled pairwise. A WRF value was not removed simply because another observation variable or the secondary station was missing. This approach preserved valid model-observation pairs while preventing infilling from creating artificial agreement. The final matched counts were therefore allowed to differ among WBGT, solar radiation, and the Tempest-derived variables.

Table 3. Observation sources and final data completeness.

Source / variable	Available	Expected	Completeness	Missing
KestrelMet/AWN WBGT	3430	3430	100.00%	0
KestrelMet/AWN natural wet bulb	3430	3430	100.00%	0
KestrelMet/AWN black globe	3430	3430	100.00%	0
KestrelMet/AWN solar radiation	3419	3430	99.68%	11
Tempest WBGT	3393	3430	98.92%	37
Tempest wet bulb	3393	3430	98.92%	37
Tempest solar radiation	3393	3430	98.92%	37

2.10 Statistical evaluation

For matched model (M) and observation (O) pairs, error was defined as $M - O$. The principal continuous statistics were bias, mean absolute error (MAE), root-mean-square error (RMSE), and Pearson correlation coefficient. Bias was used to identify systematic warm or cool errors; MAE described the typical absolute error; RMSE placed more weight on larger departures; and correlation described how well the modeled time series followed the observed temporal behavior.

$$\text{Bias} = (1/N) \sum (M_i - O_i) \quad (6)$$

$$\text{MAE} = (1/N) \sum |M_i - O_i| \quad (7)$$

$$\text{RMSE} = \sqrt{(1/N) \sum (M_i - O_i)^2} \quad (8)$$

Whole-day statistics were calculated from all matched 15-min values across the study period. Daily RMSE or MAE values were not averaged to create the project-level metrics. Daily-maximum analysis was performed separately by identifying the observed and modeled maximum for each validation day. Peak magnitude error and peak timing error were then calculated for each method and domain. This separation was necessary because a method can perform well over the full daytime period but still miss the magnitude or timing of the daily heat-stress maximum.

2.11 Statistical rationale and interpretation

No single statistical measure provides a complete description of model performance. A model can have a small average bias while still producing large individual errors: it can have a strong correlation with observations while remaining systematically too warm or too cool, and it can reproduce daytime

conditions reasonably well while missing the magnitude or timing of the daily maximum. For these reasons, we evaluated the WRF-derived WBGT calculations with several complementary statistics rather than selecting a preferred configuration from one metric alone.

Bias was used to identify systematic directional error. Error was defined as modeled minus observed. Therefore, a positive bias indicates that the model tended to produce values that were too high, while a negative bias indicates values that were too low. Bias is useful because systematic errors in temperature, moisture, wind, radiation, or a WBGT component can reveal a physical problem in the model or post-processing pathway. Bias alone, however, can be misleading. Positive and negative errors can cancel when averaged, producing a bias near zero even when individual errors are large. We therefore did not use bias by itself to rank the WBGT configurations.

Mean absolute error (MAE) was used to describe the typical magnitude of the model error without allowing positive and negative errors to cancel. MAE remains in the same units as the variable being evaluated, which makes it straightforward to interpret. For example, an MAE of 3°F means that the modeled WBGT differed from the observation by approximately 3°F on an average matched timestamp, regardless of whether the error was positive or negative. MAE is also less sensitive to a small number of unusually large errors than RMSE. Consequently, it provides a useful measure of the routine error experienced over the validation period.

Root-mean-square error (RMSE) was used as the primary overall error metric because it gives additional weight to larger departures between the modeled and observed values. Large WBGT errors are important in a heat-stress application because a relatively small number of poor forecasts can be operationally important, particularly when they occur during the hottest portion of the day. RMSE therefore complements MAE: MAE describes the typical absolute error, while RMSE is more sensitive to the larger errors in the distribution. When RMSE is substantially larger than MAE, the difference also indicates that relatively large errors occurred even if the typical error was smaller. Both metrics were considered when interpreting performance rather than assuming that the configuration with the smallest RMSE was automatically superior in every respect.

Pearson correlation coefficient (r) was used to determine how well the modeled values followed the observed temporal variations. Correlation answers a different question from MAE or RMSE. A model can have a high correlation because it reproduces the timing of warming and cooling through the day, yet still maintain a substantial warm or cool bias. Conversely, a model with a small mean error can have weaker correlation if it does not reproduce the temporal variability correctly. Correlation was therefore treated as a measure of temporal agreement rather than as a stand-alone measure of model accuracy. The study did not rank the WBGT configurations solely by correlation or coefficient of determination because those statistics do not measure absolute agreement between modeled and observed values.

The study also evaluated daily-maximum magnitude error and peak-timing error separately from the full-day statistics. This distinction was made because WBGT is often used to identify periods of greatest heat stress. A model can reproduce the average daytime WBGT reasonably well while underestimating the daily maximum, or it can reproduce the magnitude of the maximum but place it one or more hours too early or too late. Daily-peak magnitude error therefore measures whether the intensity of the highest heat stress was reproduced, while peak-timing error measures whether the maximum occurred at the correct time. These statistics were not substituted for the full-day evaluation, they were used as an additional operationally relevant test.

Percentage-error statistics, such as mean absolute percentage error (MAPE), were not used as primary metrics. WBGT and its temperature components are measured on temperature scales with an arbitrary zero point, expressing their error as a percentage can therefore be difficult to interpret physically. Percentage errors also become unstable for variables that can approach zero, such as near-surface wind speed. Absolute-error statistics expressed in the original units were more appropriate for comparing the meteorological drivers and WBGT components used in this study.

Formal inference also required consideration of the temporal structure of the dataset. The 49 observations collected between 0700 and 1900 on a given day are not statistically independent. Temperature, moisture, radiation, and WBGT at 1215 local time are strongly related to the conditions at 1200 and 1230 local time. Treating every 15-min value as an independent sample would artificially increase the effective sample size and could make very small differences between methods appear more statistically certain than they are. For this reason, inferential comparisons were based on whole-day bootstrap resampling, described in Section 2.12, rather than on an independent-observation t test or a bootstrap that resampled individual 15-min timestamps.

The bootstrap analysis was also paired when two model configurations were compared. The same resampled days were used for both methods within each replicate, and the difference between their performance metrics was calculated directly. This approach removes much of the day-to-day weather variability from the comparison because both configurations are evaluated under the same set of atmospheric conditions. A paired 95% bootstrap interval that included zero was interpreted as an unresolved difference. Therefore, a configuration was not described as statistically superior simply because its aggregate RMSE was numerically smaller.

Finally, the bootstrap's frequency, mean rank, and top-three frequency were used to describe ranking stability. Winner frequency represents the percentage of bootstrap samples in which a configuration produced the lowest RMSE among all candidates. Mean rank indicates its average position across the bootstrap samples, while top-three frequency describes how often it remained among the leading configurations. These quantities provide information that a single deterministic ranking cannot. A method

may have the lowest RMSE calculated from the complete 70-day dataset but be the top performer (← same question as above) only a modest fraction of bootstrap samples, indicating that its apparent advantage is sensitive to which days are included. Conversely, a configuration that repeatedly remains near the top of the rankings can be considered more robust even when its nominal advantage over another method is small.

Taken together, these statistics were chosen to answer different parts of the validation problem. Bias describes systematic direction; MAE describes typical error magnitude; RMSE emphasizes larger errors; correlation describes temporal coverability; daily-maximum metrics evaluate peak heat stress; and the whole-day bootstrap quantifies uncertainty in the resulting rankings. Agreement among several metrics provides stronger evidence than any one statistic alone. Where the metrics disagreed, the differences were treated as scientifically meaningful information about model behavior rather than reduced to a single overall score.

2.12 Whole-day bootstrap and ranking uncertainty

The preceding continuous statistics describe performance for the observed 70-day dataset; however, they do not by themselves indicate how sensitive a method ranking is to the particular sequence of weather conditions included in the study. Ranking uncertainty was assessed with a nonparametric whole-day bootstrap. The local validation day was used as the resampling unit so that the 49 within-day observations remained together and their temporal dependence was preserved. Each replicate sampled 70 validation days with replacement and retained all matched 15-min observations associated with the sampled days. The final analysis used 10,000,000 bootstrap replicates with seed 20260823.

Pairwise differences were used for inference. The same resampled days were applied to both configurations within each paired comparison so that RMSE, MAE, and bias differences were evaluated under identical sampled weather conditions. The study did not describe one configuration as statistically better when the paired 95% bootstrap interval for method A minus method B included zero. This decision rule was important because several leading configurations differed by only hundredths of a degree Fahrenheit in aggregate RMSE. A static ranking table can make differences of that size appear more meaningful than the resampling uncertainty supports.

2.13 Descriptive heat-risk categories

Continuous statistics were used for the primary ranking analysis. For operational context, the workflow also displayed project-defined heat-risk bands of Low (<78.3°F), Elevated (78.3-82.0°F), Moderate (82.1-86.0°F), High (86.1-90.0°F), and Extreme (>90.0°F). These categories were used descriptively to

compare classification behavior; they were not treated as a regulatory standard and were not used to select the preferred WBGT pathway.

2.14 Use of generative artificial intelligence

Generative artificial intelligence tools (OpenAI ChatGPT) were used during preparation of this manuscript to assist with code development and debugging, statistical workflow development, figure preparation, and editorial refinement of manuscript text. All model configurations, methodological decisions, data processing procedures, statistical analyses, scientific interpretations, and conclusions were reviewed and verified by the author. The author assumes full responsibility for the accuracy and integrity of the work.

3. Results

3.1 Data completeness and observed heat conditions

The final dataset contained all 70 intended validation dates and 10,290 WRF-domain rows. Ambient WBGT, natural wet bulb, and black globe were complete at all 3,430 target times. Tempest WBGT was available at 3,393 target times. The 37 missing values were confined to four partially incomplete days rather than a loss of entire validation days. No duplicate validation-date/time/domain rows were present, and all three WRF domains contained the expected 49 daytime timestamps for every date.

The KestrelMet/AWN observations show that the study period contained sustained summer heat rather than a collection of marginal WBGT days. Mean WBGT over 0700-1900 was 85.59°F, the median was 87.80°F, and the observed range was 68.0-98.8°F. The mean daily maximum WBGT was 93.0°F. These values provided a broad range of elevated, high, and extreme heat-stress conditions for testing the model-derived calculations.

Tempest reported a lower mean WBGT of 83.34°F and a mean daily maximum of 91.34°F; its overall WBGT range was 68.90-98.24°F. The two stations therefore described the same general heat events, but they did not provide the same WBGT baseline. That difference became important when the model configurations were ranked against each source.

3.2 Observation-source differences

On the 3,393 timestamps shared by both stations, KestrelMet/AWN and Tempest WBGT were strongly correlated ($r = 0.941$). However, Ambient WBGT averaged 2.274°F warmer. The source-to-source WBGT MAE was 2.541°F, and the RMSE was 3.221°F. The magnitude of this offset is important because it is much larger than the RMSE difference between several of the leading WRF-derived configurations.

The WBGT difference also changed through the day. Mean Ambient-minus-Tempest WBGT was only about 0.16°F at 0700 and 0.64°F at 0800. It increased to approximately 2.78°F at 0900 and 3.84°F at 1000, remained generally 2-3°F through much of the afternoon, and then decreased late in the validation period. This pattern indicates that the source difference is not a simple constant offset that could be removed with one correction.

The standard meteorological variables agreed more closely than WBGT. Air temperature differed by 0.821°C on average with $r = 0.991$; station pressure differed by 0.793 hPa with $r = 0.999$; and dew point differed by -0.406°C. Wind and wet-bulb-related quantities showed larger source differences. This matters because relatively small differences in several meteorological inputs can combine into a larger difference in the final WBGT.

Table 4. Direct comparison of the two observation sources on common timestamps. Difference is KestrelMet/AWN minus Tempest.

Variable	N	Ambient means	Tempest means	Mean diff.	RMSE	r	Units
WBGT	3393	85.613	83.339	+2.274	3.221	0.941	°F
Air temperature	3393	31.555	30.734	+0.821	1.071	0.991	°C
Dew point	3393	22.198	22.604	-0.406	0.840	0.962	°C
Wind speed	3393	0.638	0.967	-0.329	0.536	0.742	m s ⁻¹
Solar radiation	3382	475.041	473.064	+1.977	139.235	0.921	W m ⁻²
Station pressure	3393	987.013	986.220	+0.793	0.828	0.999	hPa
Wet-bulb temperature	3393	26.728	25.050	+1.678	2.129	0.826	°C

3.3 Base-pathway performance

Before the Carter/Bernard and FAO branches were applied, the three pathways behaved very differently. Against Ambient, Liljegren Base produced the lowest Base-pathway RMSE in each domain: 5.088°F on d01, 5.130°F on d02, and 4.993°F on d03. Dimiceli/Piltz Base was substantially cooler, while Hunter Base was strongly warm. On d01, Hunter Base had a WBGT bias of +11.76°F and an RMSE of 13.90°F; the same general behavior occurred in d02 and d03.

The Base-pathway ordering was similar against Tempest, but the offsets changed. Liljegren Base RMSE ranged from 3.200°F on d01 to 3.348°F on d03, with biases from -0.986 to -0.552°F. Dimiceli/Piltz Base remained cooler than the observations, and Hunter Base remained strongly warm.

This confirms that the large differences among the unmodified pathways were not caused by one observation source alone.

Table 5. Base-pathway full-day performance for d01. Errors are model minus observation, in degrees Fahrenheit.

Source	Pathway	Bias	MAE	RMSE	r
Ambient	Hunter	+11.757	11.922	13.898	0.826
Ambient	Dimiceli/Piltz	-5.767	6.255	7.243	0.792
Ambient	Liljegren	-3.246	4.297	5.088	0.813
Tempest	Hunter	+14.022	14.141	16.119	0.817
Tempest	Dimiceli/Piltz	-3.511	4.068	4.849	0.826
Tempest	Liljegren	-0.986	2.456	3.200	0.836

3.4 Effects of Carter/Bernard and FAO branches

The largest branch improvement occurred in the Hunter family. Against Ambient d01, Hunter Base RMSE was 13.898°F, while Hunter Carter_FAO reduced it to 3.949°F. Bias changed from +11.757°F to -0.965°F. The improvement came primarily from replacing the Hunter Base Tn treatment rather than from changing the final WBGT weighting.

Carter substitution also improved the Liljegren family against Ambient. On d01, Liljegren Base RMSE was 5.088°F, Liljegren Carter was lower, and Liljegren Carter_FAO reached 3.914°F. The FAO branch by itself did not provide the same improvement in every case, its effect depended on whether reducing the wind input corrected or amplified errors already present in Tn and Tg.

Dimiceli/Piltz remained less competitive as an end-to-end WBGT pathway. Its Tg estimates were generally cool, and the Base/FAO branches used psychrometric wet bulb as the Tn proxy. Carter substitution improved the final WBGT, especially when combined with FAO, but the pathway still did not match the full-day performance of the leading Liljegren and Hunter hybrid branches. The full-day ranking summaries, the method-family comparison, and the branch comparison are shown in Figures 3 through 6.

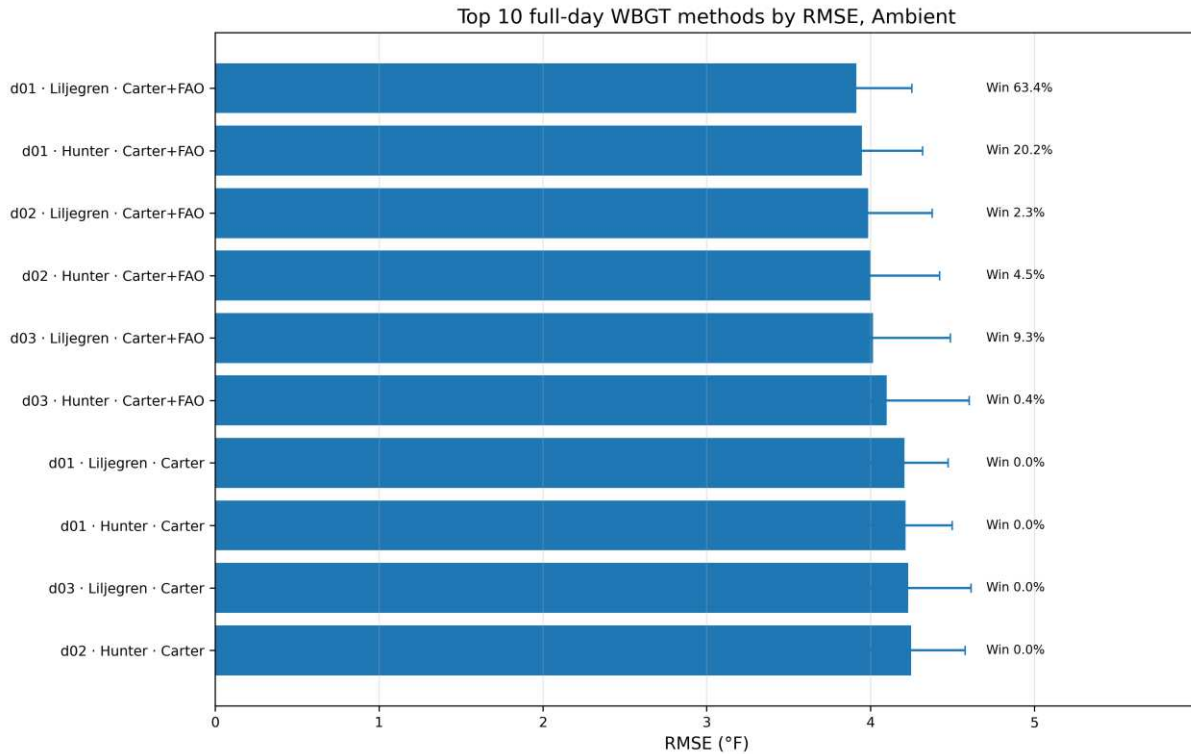


Fig. 3 Top 10 full-day WBGT methods by RMSE against KestrelMet/AWN observations. Error bars denote the 2.5th to 97.5th percentile bootstrap RMSE range, and right-side labels show the bootstrap winner frequency

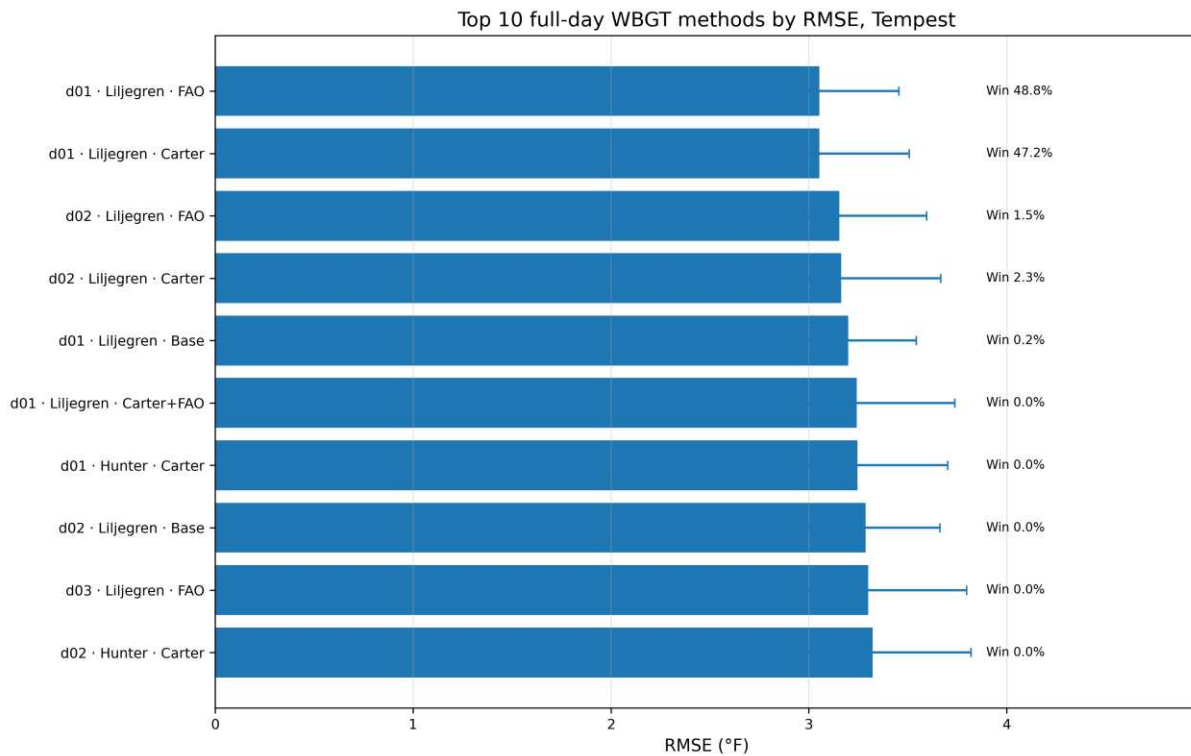


Fig. 4 Top 10 full-day WBGT methods by RMSE against Tempest observations. Error bars denote the 2.5th to 97.5th percentile bootstrap RMSE range, and right-side labels show the bootstrap winner frequency

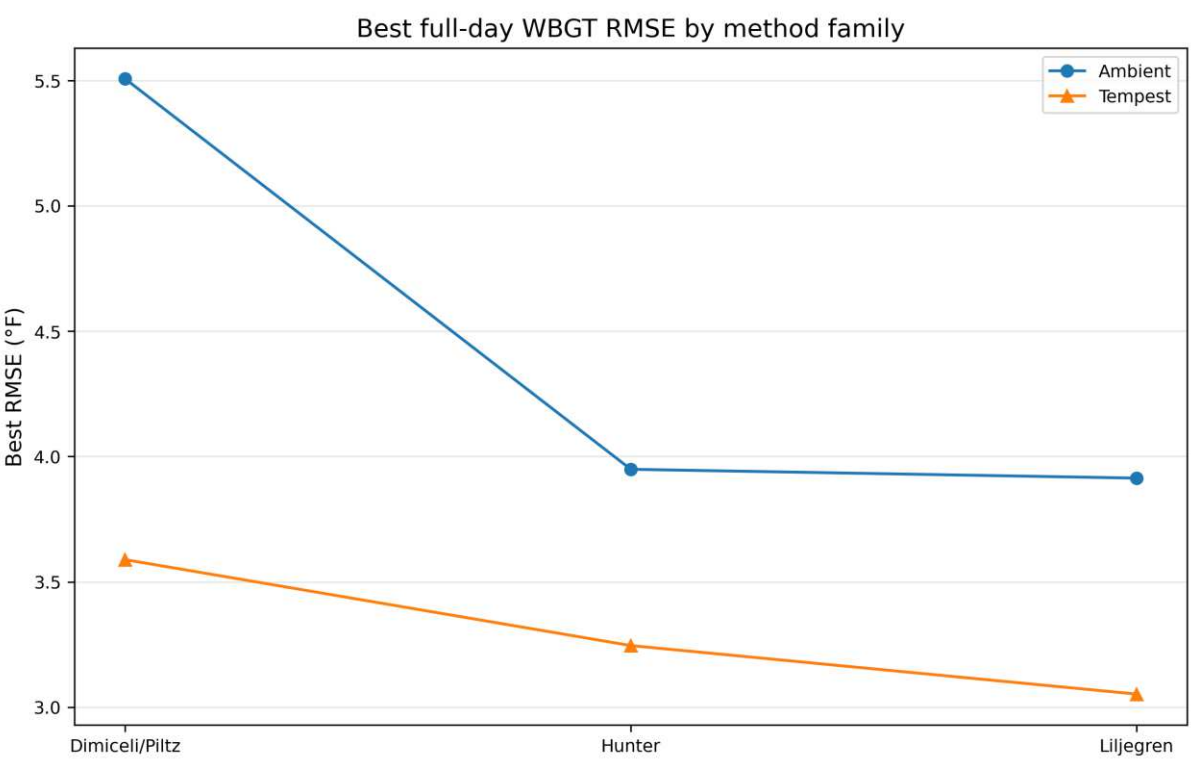


Fig. 5 Best full-day WBGT RMSE by method family for KestrelMet/AWN and Tempest validation sources. Lower values indicate better agreement with the corresponding observation source

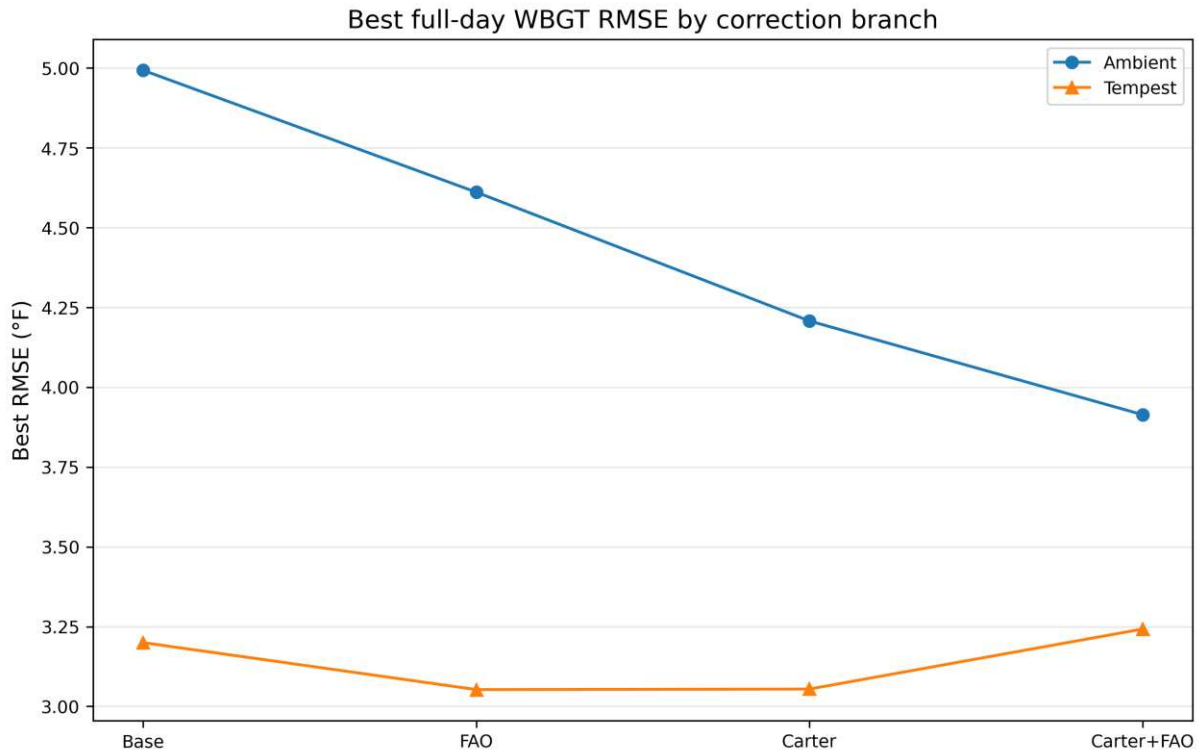


Fig. 6 Best full-day WBGT RMSE by correction branch for KestrelMet/AWN and Tempest validation sources. The branch comparison highlights the impact of the Base, FAO, Carter, and Carter_FAO treatments on final WBGT performance

3.5 Grid-spacing sensitivity

Full-day performance did not improve steadily as grid spacing decreased. For Ambient Liljegren Carter_FAO, RMSE increased from 3.914°F on d01 to 3.986°F on d02 and 4.015°F on d03. Hunter Carter_FAO showed a similar full-day pattern. Against Tempest, d01 also produced the strongest full-day Liljegren results. In other words, the 25-km domain was not simply outperformed by the 5-km and 1-km nests for daytime aggregate WBGT as would be expected by high resolution numerical weather modeling

Daily maxima produced a different pattern. Ambient peak RMSE was lowest for d03 Liljegren Carter_FAO at 4.001°F, while d03 Hunter Carter_FAO was nearly identical at 4.007°F. For Tempest peaks, d01 Hunter Carter_FAO (3.597272°F) and d03 Liljegren Carter_FAO (3.597476°F) were effectively tied in the aggregate statistics. The preferred grid spacing therefore depended on whether the target was full-day behavior or the daily maximum.

This result is reasonable for a point validation. Finer resolution can sharpen convection, cloud timing, land-surface gradients, and wind structure, but it does not guarantee that those features occur at the correct place and time. A 1-km WRF grid also cannot reproduce the immediate sheltering around a residential instrument site. The grid-spacing results should therefore be interpreted as a sensitivity of the

complete model-to-point comparison, not as proof that one horizontal resolution is universally better. The best full-day RMSE attained within each domain is summarized in Figure 7.

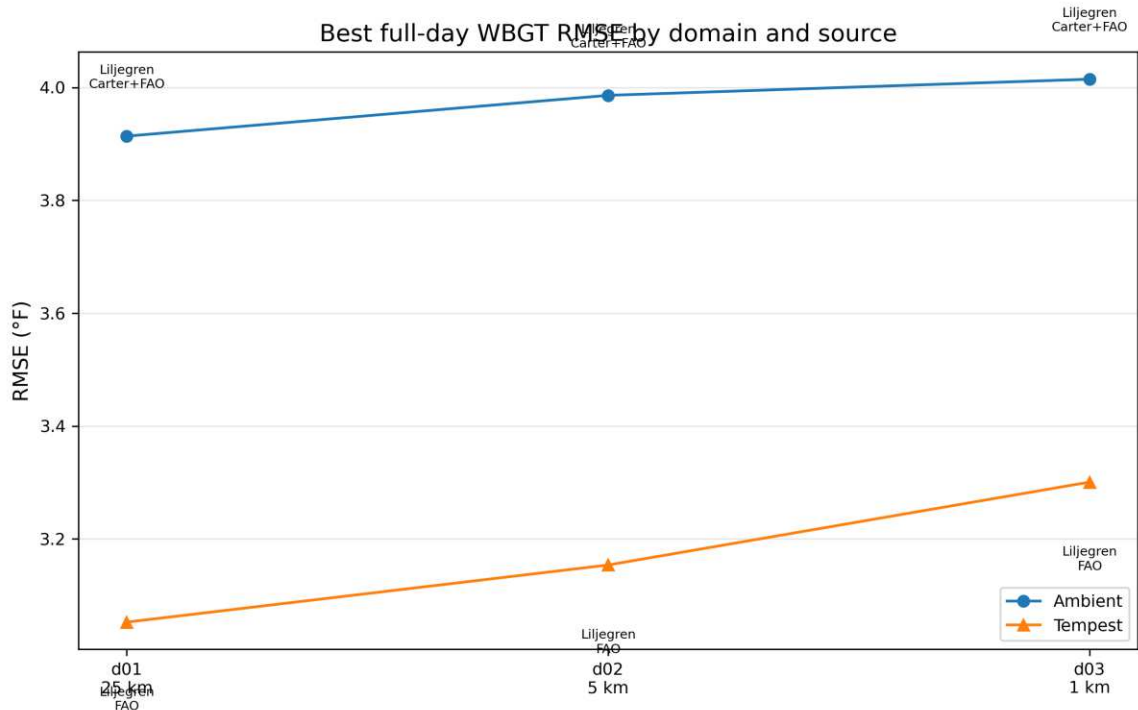


Fig. 7 Best full-day WBGT RMSE attained within each domain for KestrelMet/AWN and Tempest validation sources. For each domain, the plotted value represents the lowest full-day RMSE found among all pathway and branch combinations

3.6 Component diagnostics: natural wet bulb and globe temperature

The component analysis explains much of the branch behavior. Ambient observed natural wet-bulb temperature averaged 80.09°F. On d01, Hunter Base Tn had a +16.515°F bias and a 19.171°F RMSE. Replacing that Tn treatment with Carter_FAO reduced the bias to about -2.3°F and the RMSE to approximately 4.0°F. Liljegren Tn was cool in the Base branch, while the Carter and Carter_FAO substitutions moved it closer to the Ambient observations.

Globe-temperature errors followed a different pattern. On d01, Liljegren Base/Carter Tg had a small full-day bias relative to Ambient and an RMSE near 7.9°F; Hunter Base/Carter Tg had a similarly small mean bias with a somewhat larger RMSE. Dimiceli/Piltz Tg was systematically cool. The expected branch structure was also preserved: Carter did not change Tg relative to Base, and Carter_FAO did not change Tg relative to FAO within the same pathway.

These component results show why a good final WBGT score must be interpreted carefully. A cool Tn can be partly offset by a warm Tg, a warm air-temperature contribution can be offset by dry moisture forcing, and a wind correction can improve one component while worsening another. Final WBGT error

alone does not show whether the underlying meteorology and heat-transfer terms are physically reasonable. Component-focused summaries for natural wet bulb and globe temperature are shown in Figures 8 and 9.

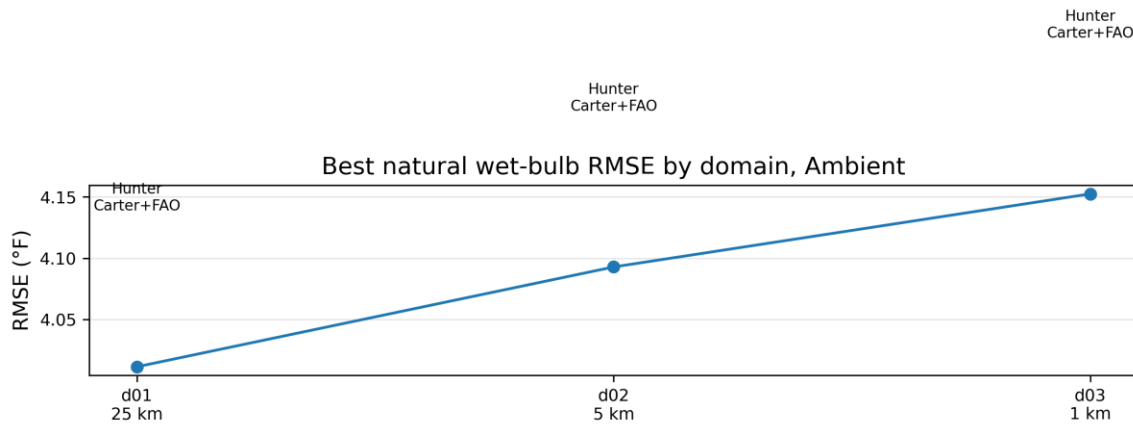


Fig. 8 Best natural wet-bulb RMSE attained within each domain using KestrelMet/AWN observations. These values summarize how well the different post-processing pathways reproduced the observed natural wet-bulb component

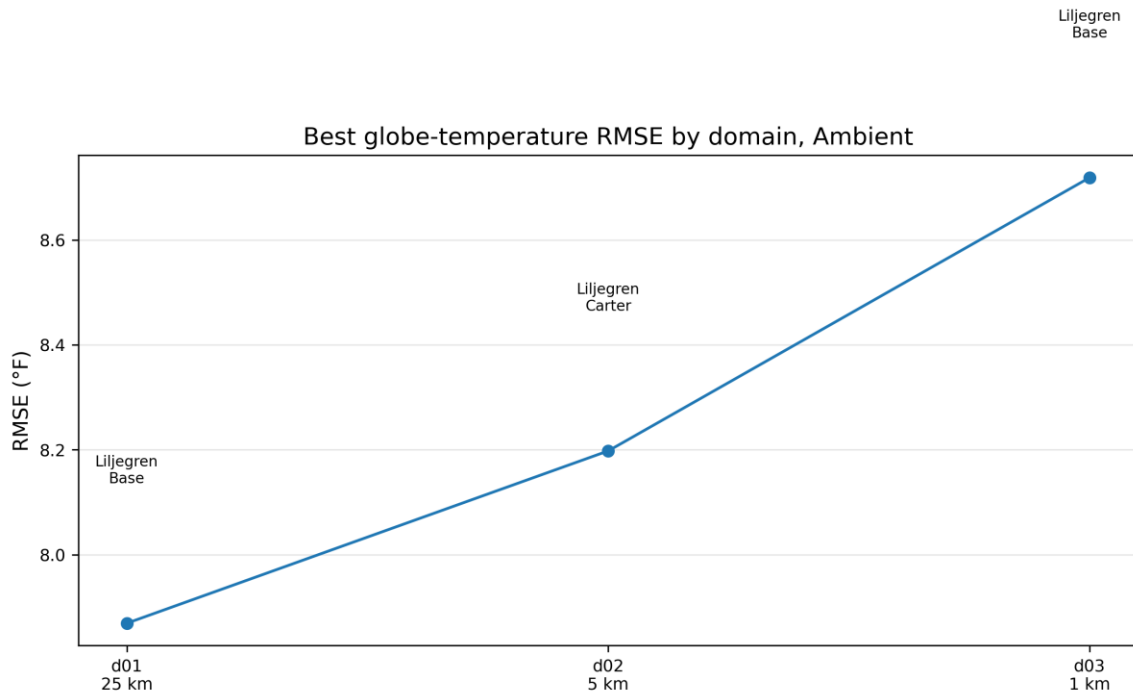


Fig. 9 Best black-globe-temperature RMSE attained within each domain using KestrelMet/AWN observations. Lower RMSE indicates better agreement with the observed globe-temperature component

3.7 Meteorological-driver errors

Air temperature was one of the better-predicted meteorological drivers, especially in d01. Against Ambient, d01 air-temperature bias was near zero, with RMSE about 1.7°C and correlation above 0.93. Bias became warmer in the finer domains. Tempest showed the same general pattern but with a warmer model offset because Tempest air temperature was cooler than Ambient on average.

Moisture showed a consistent dry bias. WRF dew point was too low against both observation sources, and the dry bias generally increased from d01 to d03. The psychrometric wet-bulb calculation was also too cool relative to the Ambient natural-wet-bulb observation. These moisture errors are important because T_n carries 70% of the outdoor WBGT weighting.

Wind was the largest persistent driver mismatch. Ambient mean observed wind was only about 0.64 m s⁻¹, while WRF 10-m wind was several meters per second stronger. The FAO 10-m-to-2-m conversion reduced the wind bias substantially. However, the adjusted wind still remained too strong at this sheltered point. Tempest measured somewhat higher wind than Ambient, but the WRF wind bias remained large against that source as well.

Shortwave radiation also showed a positive WRF bias on the matched full-day records, while station pressure was modeled well. Given that pressure is a synoptic weather feature the lateral boundary conditions from the input data does not change much. The combination of dry moisture, excessive wind, and excessive solar forcing is important because these errors act in different directions within the WBGT calculation. A dry or windy bias tends to cool wet-bulb-related terms; excessive solar radiation tends to warm the radiative components. This creates the possibility of compensating error in the final WBGT. Meteorological-driver RMSE summaries are provided in Figures 10 and 11.

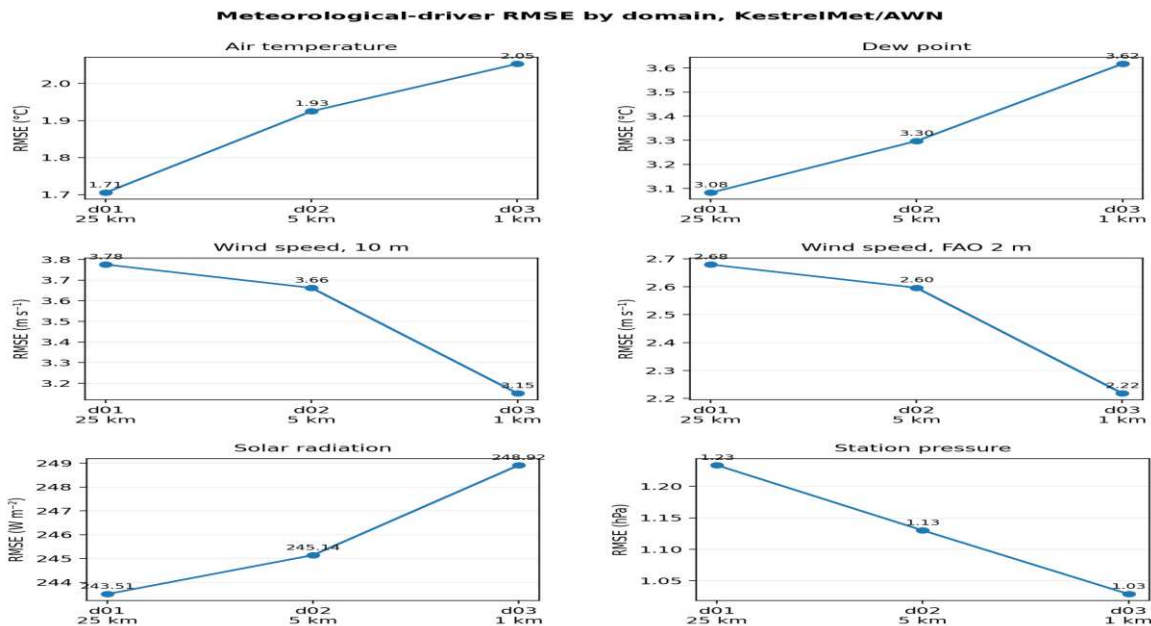


Fig. 10 Meteorological-driver RMSE by domain for validation against KestrelMet/AWN observations. Each panel uses the native units of the indicated variable so that air temperature, dew point, wind, solar radiation, and station pressure can be interpreted without mixing unlike units on one axis

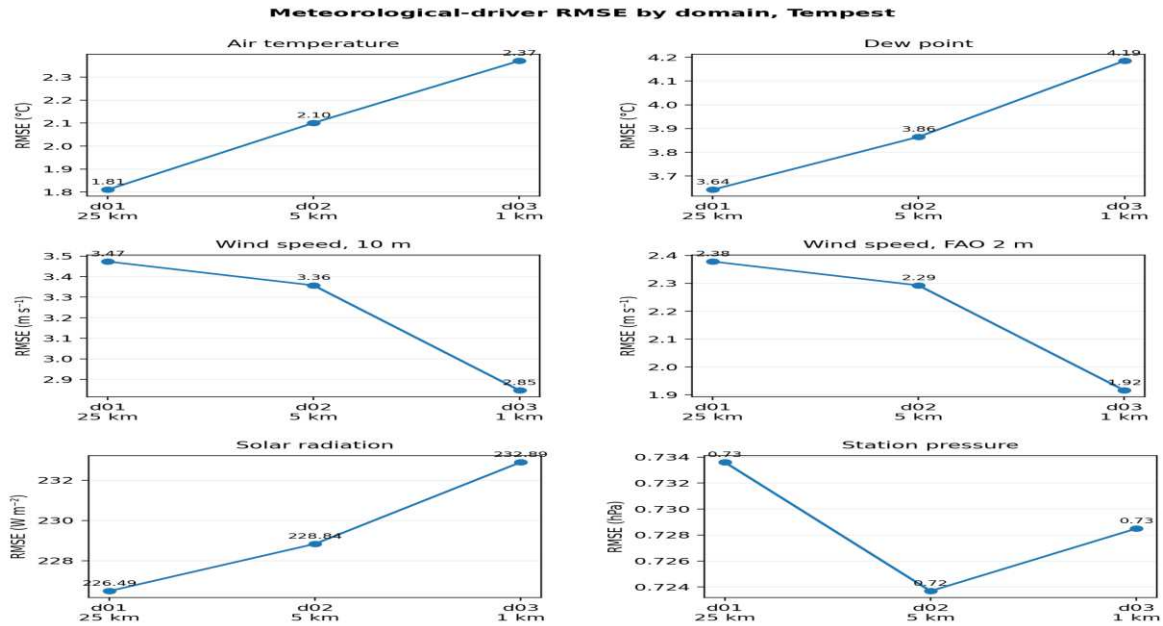


Fig. 11 Meteorological-driver RMSE by domain for validation against Tempest observations. Each panel uses the native units of the indicated variable; the panel structure matches Figure 10 for direct source-to-source comparison

Table 6. Selected d01 meteorological-driver errors. Bias is WRF minus observation.

Source	Driver	Bias	RMSE	r	Units
Ambient	air temperature	-0.037	1.705	0.938	°C
Ambient	dew point	-2.376	3.083	0.789	°C
Ambient	wind speed, 10 m	+3.506	3.775	0.400	m s ⁻¹
Ambient	wind speed, FAO 2 m	+2.462	2.679	0.401	m s ⁻¹
Ambient	solar radiation	+126.057	243.509	0.816	W m ⁻²
Ambient	station pressure	-1.073	1.233	0.983	hPa
Tempest	air temperature	+0.787	1.811	0.942	°C
Tempest	dew point	-2.786	3.642	0.688	°C
Tempest	wind speed, 10 m	+3.180	3.473	0.415	m s ⁻¹
Tempest	wind speed, FAO 2 m	+2.135	2.378	0.415	m s ⁻¹
Tempest	solar radiation	+127.592	226.494	0.841	W m ⁻²
Tempest	station pressure	-0.278	0.734	0.982	hPa

3.8 Daily-maximum magnitude and timing

Daily-maximum magnitude did not produce the same ranking as the full-day analysis. Against Ambient, d03 Liljegren Carter_FAO had the nominally lowest peak RMSE at 4.001°F; d03 Hunter Carter_FAO was only 0.0055°F higher and had the lower MAE and smaller absolute signed bias. Against Tempest, d01 Hunter Carter_FAO and d03 Liljegren Carter_FAO had virtually identical peak RMSE values. These differences are too small to interpret from the static metric table alone.

Peak timing remained a clear limitation. The leading Ambient peak configurations had mean absolute timing errors on the order of 90-100 min, and the Tempest peak configurations were similar. A method can therefore estimate the daily maximum magnitude reasonably well and still place that maximum more than an hour away from the observed time. That limitation matters for applications that depend on the timing of activity restrictions or peak heat exposure. Daily-peak ranking summaries, the best daily-peak RMSE by domain, and the hourly evolution of WBGT RMSE are shown in Figures 12 through 16.

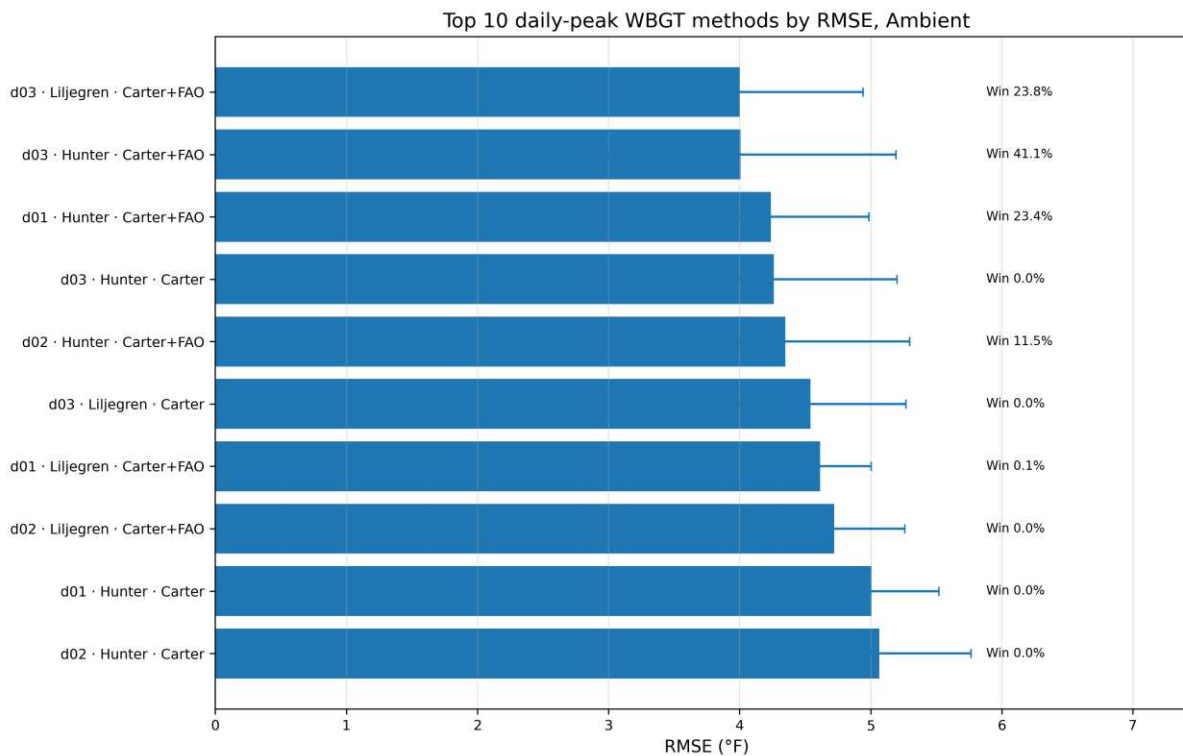


Fig. 12 Top 10 daily-peak WBGT methods by RMSE against KestrelMet/AWN observations. Error bars denote the 2.5th to 97.5th percentile bootstrap RMSE range, and right-side labels show the bootstrap winner frequency

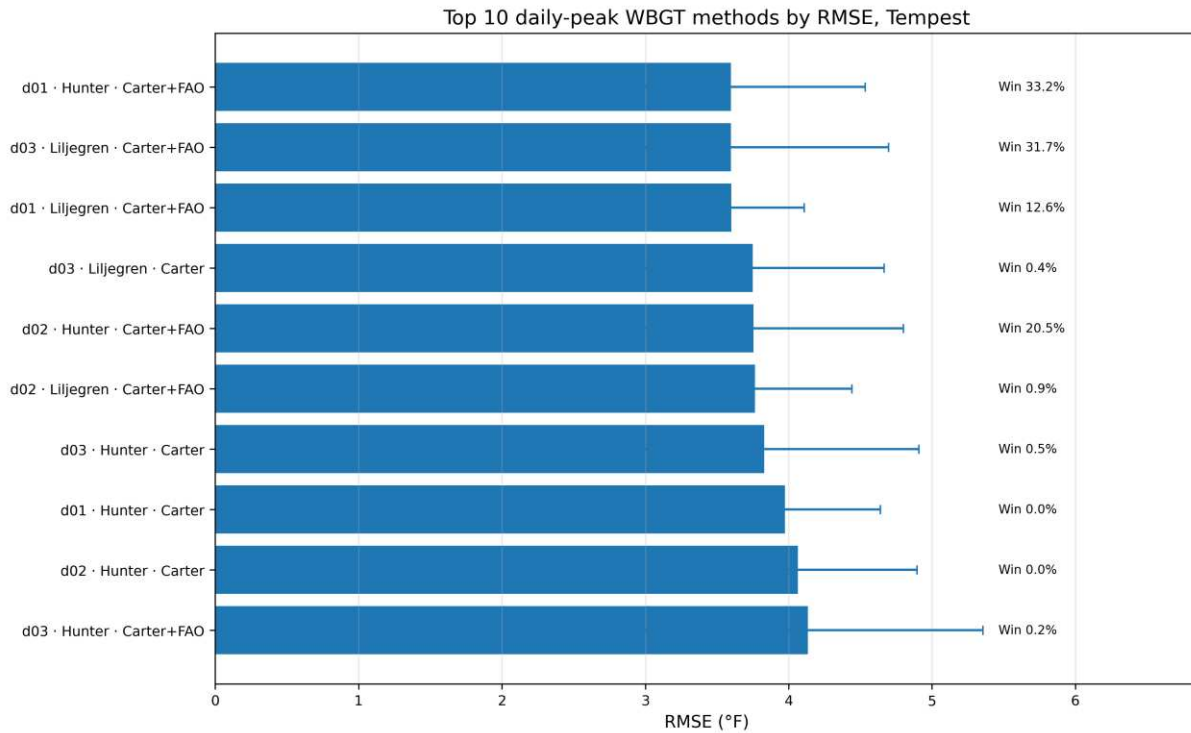


Fig. 13 Top 10 daily-peak WBGT methods by RMSE against Tempest observations. Error bars denote the 2.5th to 97.5th percentile bootstrap RMSE range, and right-side labels show the bootstrap winner frequency

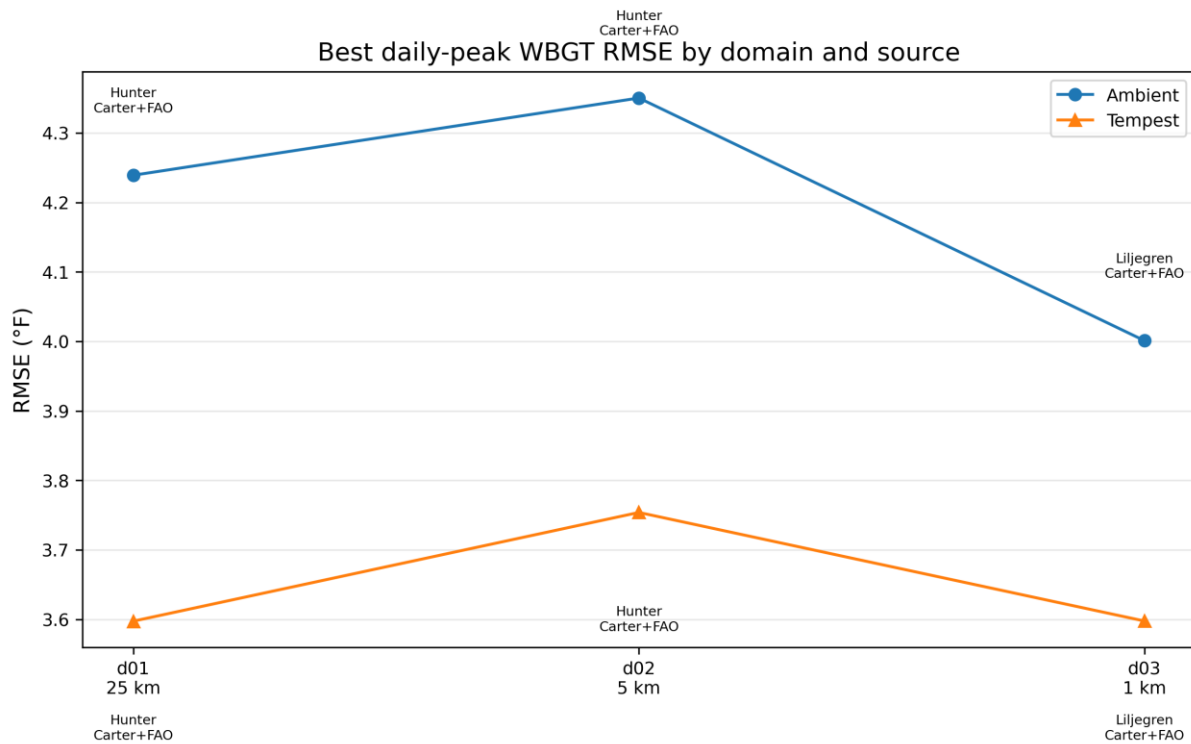


Fig. 14 Best daily-peak WBGT RMSE attained within each domain for KestrelMet/AWN and Tempest validation sources. For each domain, the plotted value represents the lowest daily-peak RMSE found among all pathway and branch combinations

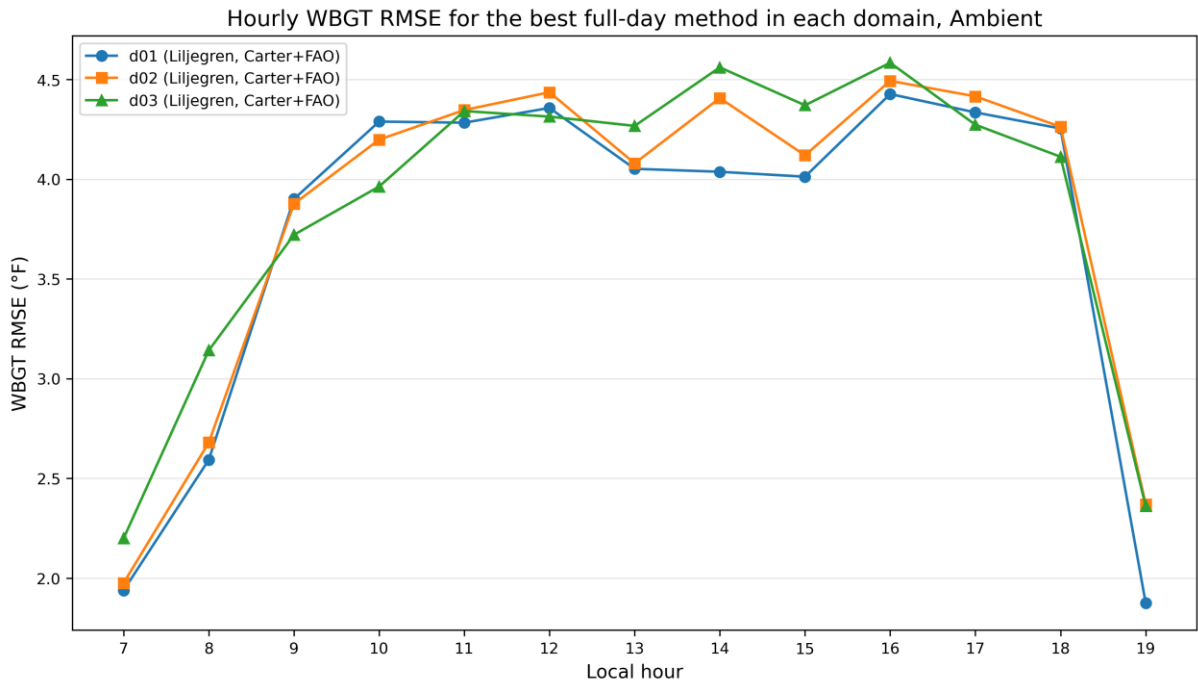


Fig. 15 Hourly WBGT RMSE from 0700 to 1900 local time for the best full-day method in each domain, using KestrelMet/AWN observations. The plot shows how error evolved through the diurnal validation window

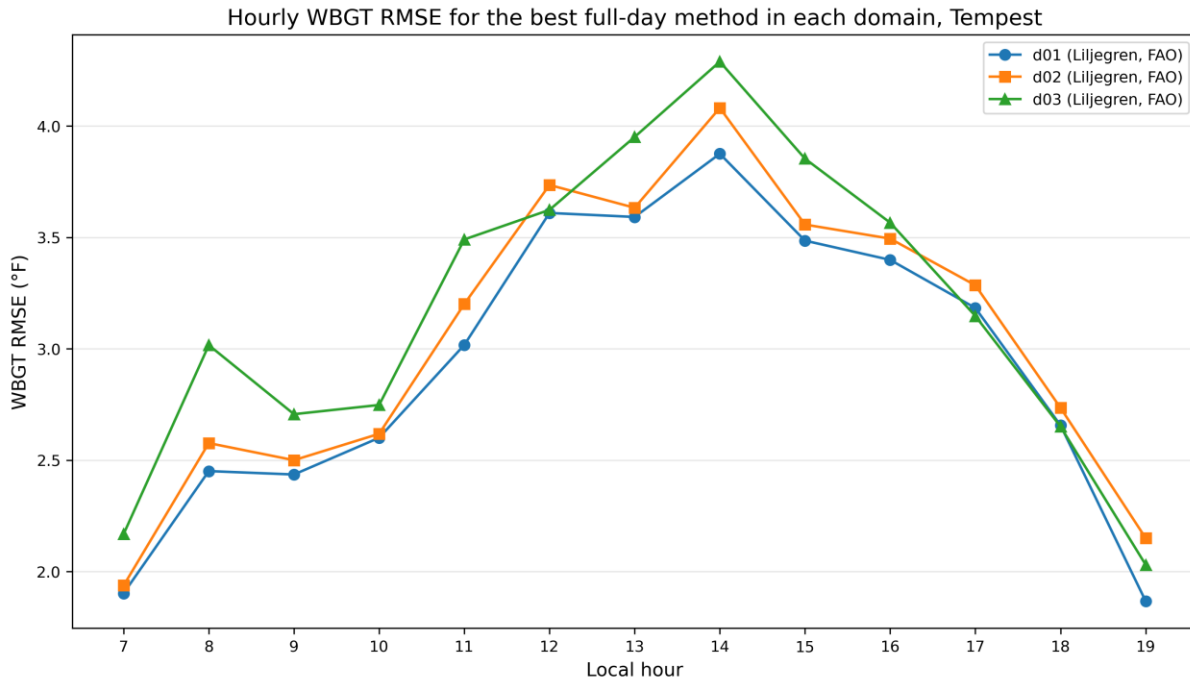


Fig. 16 Hourly WBGT RMSE from 0700 to 1900 local time for the best full-day method in each domain, using Tempest observations. The plot shows how error evolved through the diurnal validation window

3.9 Bootstrap ranking robustness

The 10,000,000-replicate whole-day bootstrap changed how the nominal rankings should be described. For Ambient full-day WBGT, d01 Liljegren Carter_FAO had an RMSE of 3.9138°F and performed the best of the RMSE ranking in 63.42% of replicates. d01 Hunter Carter_FAO had an RMSE of 3.9490°F. The paired RMSE difference was -0.0352°F, with a 95% bootstrap interval from -0.1307 to +0.0534°F; the interval crossed zero, so the difference was not statistically resolved.

Tempest full-day results were even closer. d01 Liljegren FAO and d01 Liljegren Carter had RMSE values of 3.0526°F and 3.0543°F, respectively. Their bootstrap winner frequencies were approximately 48.8% and 47.2%, and the paired RMSE interval crossed zero. These two branches should therefore be described as statistically indistinguishable co-leaders rather than ranked as first and second in a meaningful statistical sense.

Peak rankings were also unresolved. For Ambient, the paired RMSE interval comparing d03 Liljegren Carter_FAO with d03 Hunter Carter_FAO was -0.421 to +0.461°F. For Tempest, the paired interval comparing d01 Hunter Carter_FAO with d03 Liljegren Carter_FAO was -1.134 to +1.257°F. The bootstrap confirms that the apparent peak winners are better described as co-leading or highly competitive configurations.

Table 7. Final headline RMSE comparisons from the 10,000,000-replicate whole-day bootstrap. Delta is method A minus method B.

Analysis	Method A	RMSE A	Method B	RMSE B	Delta	95% paired CI	P(A lower)	Resolved?
Ambient full day	d01 Liljegren Carter_FAO	3.914	d01 Hunter Carter_FAO	3.949	- 0.035	[-0.131, +0.053]	76.3%	No
Tempest full day	d01 Liljegren FAO	3.053	d01 Liljegren Carter	3.054	- 0.002	[-0.099, +0.103]	50.4%	No
Ambient daily max	d03 Liljegren Carter_FAO	4.001	d03 Hunter Carter_FAO	4.007	- 0.006	[-0.421, +0.461]	48.6%	No
Tempest daily max	d01 Hunter Carter_FAO	3.597	d03 Liljegren Carter_FAO	3.597	- 0.000	[-1.134, +1.257]	50.5%	No

3.10 Descriptive heat-risk threshold behavior

The difference between observation sources also changed the frequency of descriptive heat-risk categories. Using the project display bands, Ambient classified 32.86% of valid timestamps as Extreme and 25.83% as High. Tempest classified fewer timestamps in the highest category because its WBGT values were lower on average. This source dependence should be kept in mind when category agreement is used for operational interpretation.

For the leading full-day configurations, exact category agreement was 38.4% for Ambient d01 Liljegren Carter_FAO and 41.2% for Ambient d01 Hunter Carter_FAO; both were within one adjacent category for approximately 90% of timestamps. Category agreement provides useful operational context, but the continuous statistics and paired bootstrap remain the primary basis for comparing the methods.

Table 8. Descriptive heat-risk category agreement for leading full-day configurations.

Source	Candidate	N	Exact	Within +/- 1	Model higher	Model lower
Ambient	d01 Liljegren Carter_FAO	3430	38.4%	90.3%	16.2%	45.4%
Ambient	d01 Hunter Carter_FAO	3430	41.2%	90.0%	17.4%	41.4%

Source	Candidate	N	Exact	Within +/- 1	Model higher	Model lower
Tempest	d01 Liljegren FAO	3393	56.9%	95.3%	16.8%	26.3%
Tempest	d01 Liljegren Carter	3393	59.7%	95.2%	23.5%	16.8%

4. Discussion

4.1 Principal findings

The most important result from this experiment is that there is no single WBGT equation, branch, or WRF grid spacing that wins every comparison (or provides the most accurate results that can be used for microclimates). Performance changed with the component pathway, the correction branch, the WRF domain, the observation source, and whether the target was full-day behavior or the daily maximum. Liljegren-based configurations were the strongest and most stable for full-day WBGT. Hunter/Minyard became competitive when its natural wet-bulb treatment was replaced with Carter/Bernard, and Hunter Carter_FAO was especially competitive for daily maxima. Dimiceli/Piltz remained useful as a simplified Tg pathway, but it did not provide the strongest complete WBGT estimates. The bootstrap is important because several differences that appear to be rankings in a static table were too small, relative to day-to-day variability, to support a claim that one method was statistically superior.

These results support a component-aware approach rather than treating WBGT as one black-box value. The final WBGT can look reasonable even when one or more meteorological drivers are biased because Tn, Tg, and Ta can offset one another. For that reason, the study considers the component outputs to be part of the validation and not an optional diagnostic. This becomes even more important if the workflow is transferred to another site, climate, or observing network.

4.2 Why Liljegren was the most robust full-day framework

Liljegren provided the strongest general full-day framework because it explicitly represents the heat and mass transfer processes that control the wetted wick and black globe. It also changes the direct/diffuse solar partition with time rather than holding it fixed. That does not mean the Liljegren Base branch was perfect. Against Ambient, its Tn remained too cool, while WRF itself was too dry, too windy, and too high in shortwave radiation. Even with those driver errors, the Liljegren family retained strong correlation and, after the Tn branch was adjusted, produced the most stable aggregate RMSE ranking.

It is important not to overstate the ranking. Ambient d01 Liljegren Carter_FAO had the highest percentage, with about 63% of the full-day RMSE bootstrap replicates, but its paired RMSE advantage over Hunter Carter_FAO was not resolved at the 95% level. Against Tempest, the leading Liljegren FAO

and Carter branches were even more clearly tied. Due to this, Liljegren is described as the preferred general framework because of its physical basis, component behavior, and stable full-day performance. The study does not claim that every Liljegren correction branch is statistically better than the closest Hunter alternative.

4.3 Carter/Bernard substitution and the dominant Tn term

The Carter/Bernard branch produced one of the largest changes in the entire experiment because it modifies Tn, which contributes 70% of outdoor WBGT. Hunter Base Tn was strongly warm in this dataset; on Ambient d01, its Tn bias was +16.5°F, and the final WBGT bias was +11.8°F. Carter_FAO reduced the Tn error dramatically and brought total WBGT bias to about -1.0°F. This explains why Hunter moved from one of the poorest Base configurations to one of the leading corrected configurations.

This result should not be read as proof that the original Hunter approach is invalid. The Hunter/Minyard relationship is empirical, and its performance depends on the meteorological inputs supplied to it. In this experiment, the combination of WRF solar radiation, wind, moisture, and the sheltered validation exposure was a poor match for the unmodified Hunter Tn regression. Carter substitution also improved Liljegren Tn against Ambient. However, that creates a hybrid pathway because the original physically based Liljegren Tn is being replaced. For operational work at this site, the hybrid performed well. For broader transferability, it is useful to retain both the original Liljegren calculation and the Carter sensitivity.

4.4 FAO wind-height adjustment: useful but not a complete shelter correction

The FAO adjustment reduced the large WRF wind bias by lowering the 10-m model wind toward an approximate 2-m exposure wind. It did not remove the wind mismatch. Even after adjustment, modeled wind remained substantially stronger than the observations, especially at the sheltered Ambient site. This is expected because FAO-56 corrects for measurement height over short grass; it does not account for a wood fence, nearby trees, a house, or other obstacles that reduce local airflow.

The effect of FAO on WBGT was therefore mixed. Lower wind reduces convective and evaporative cooling, which can warm both Tn and Tg. That change helps when the unadjusted pathway is too cool, but it can hurt when another part of the calculation is already too warm. In this study, Carter_FAO worked well for Ambient and for Hunter peak performance, while Tempest full-day results placed Liljegren FAO and Liljegren Carter in a statistical tie. The practical conclusion is to retain FAO as a physically defensible height sensitivity, but not to treat it as a universal shelter correction.

4.5 Dimiceli/Piltz as a Tg pathway

The Dimiceli/Piltz results are easiest to understand when the formulation is described correctly. Dimiceli/Piltz supplies a black-globe-temperature estimate, but it does not provide a complete independent Tn pathway. In this workflow, its Tg was paired with either psychrometric wet bulb or Carter/Bernard Tn to produce complete WBGT. The Tg values were generally cool, which explains much of the cool WBGT bias in the Base and FAO branches. Reducing wind with FAO warmed the globe calculation, and Carter improved the Tn side of the equation. Even so, the combined pathway did not match the full-day or peak performance of the leading Liljegren and Hunter configurations.

4.6 Grid spacing and point representativeness

One of the more practical findings is that finer horizontal resolution did not automatically improve point WBGT. It is easy to assume that a 1-km nest should outperform a 25-km domain, but WBGT depends on several fields that respond differently to resolution. A finer grid can sharpen cloud boundaries, convective timing, terrain effects, land-surface gradients, and wind structure. If those features are displaced in time or space, the point error can increase rather than decrease. In this study, d01 performed best for the strongest full-day configurations, while d03 became competitive for daily maxima.

The sheltered San Marcos site adds another layer to that comparison. Observed wind was much lower than WRF wind, and nearby obstacles are reasonable contributors to those lower readings. At the same time, this paper does not assign every wind or radiation error to sheltering; some portion can come from WRF physics, synoptic forcing, cloud timing, or the difference between a grid-cell average and a point observation. The present study quantifies the mismatch. A future microclimate paper could examine the local exposure in more detail.

4.7 Observation-source uncertainty is part of the validation problem

KestrelMet/AWN and Tempest followed the same general heat events, but their mean WBGT difference of about 2.27°F was much larger than the RMSE difference between many of the leading model configurations. That difference changed which branch appeared best. Ambient favored Carter_FAO configurations; Tempest full-day results favored Liljegren FAO and Carter. The two sources also differed in wet-bulb temperature, wind, and the diurnal WBGT amplitude. A validation study that reports a single 'best' model configuration without naming the observation source can therefore give a misleading impression of certainty.

This is consistent with the broader WBGT literature. Instrument design, exposure, local wind, radiation, and the method used to derive secondary variables can all affect the reported WBGT and its components (Carter et al. 2020; Clark and Konrad 2024). Keeping the two observation sources separate

was therefore a deliberate part of the study design. It allowed model uncertainty and observation-source uncertainty to be compared directly instead of hiding both inside one pooled reference dataset.

4.8 Compensating errors and physical interpretation

The driver and component results also show why using final WBGT readings alone can be misleading. WRF air temperature was comparatively accurate in d01, but dew point was too low, wind was much too strong, and shortwave radiation was too high. A dry bias tends to lower wet-bulb-related terms; excessive wind increases convective and evaporative cooling; and excessive solar radiation increases radiative heating. These errors can offset one another. The same issue appears within the correction branches: FAO can reduce a wind bias while warming Tg at the same time.

For that reason, the preferred configuration should not be selected from RMSE alone. MAE, signed bias, correlation, daily-maximum behavior, peak timing, Tn and Tg performance, and the physical meaning of the meteorological drivers all add information. The bootstrap leads to the same practical conclusion: when two candidates differ by only hundredths of a degree Fahrenheit in RMSE and the paired interval crosses zero, the component behavior is more informative than the nominal rank.

4.9 Operational implications

For an operational WRF-derived WBGT product in central Texas, these results support Liljegren as the default physically based framework for full-day conditions. At a sheltered point like the San Marcos site, Carter and Carter_FAO should remain available as sensitivity branches because they can materially improve Tn and the final WBGT. Hunter Carter_FAO is also a credible operational alternative, particularly when the daily maximum is the main target. Dimiceli/Piltz is best retained as a simplified Tg comparison and diagnostic rather than the primary end-to-end recommendation.

In practice, reporting more than one well-performing configuration may be more useful than presenting a single value with false precision. A preferred Liljegren estimate can be shown alongside a closely competing Hunter Carter_FAO estimate, with the meteorological-driver diagnostics and observation-source context available for interpretation. The bootstrap results show why this is a reasonable practice. In the study, after 70 days several leading differences were still not statistically resolved.

4.10 Limitations

- This experiment used one geographic validation site and one summer season. The results should be tested in additional climates, land-cover regimes, and synoptic conditions before they are treated as universal recommendations.

- The validation site was a sheltered residential exposure rather than an open-field reference site. This is a realistic outdoor heat-exposure environment, but it limits direct attribution of wind and radiation mismatch to WRF physics alone.
- Tempest WBGT and wet-bulb quantities are derived products and are not equivalent to dedicated WBGT component measurements, they were intentionally retained as a secondary source rather than pooled with the KestrelMet/AWN record.
- The FAO-56 adjustment is a wind-height conversion over short grass, it does not explicitly represent obstacle sheltering, site-specific roughness, or house-scale turbulence.
- Only one WRF physics configuration and one initialization strategy were tested. WRFDA/3DVAR, alternate PBL, radiation, or cumulus options, and ensemble uncertainty are outside the scope of Paper 1.
- Peak-magnitude skill was better than peak-timing skill. Mean absolute peak-timing errors near 1.5 h for the leading configurations limit applications that require precise timing of the daily heat-stress maximum.
- The descriptive threshold analysis depends on the project category cut points and rounding convention. Continuous metrics and paired whole-day bootstrap comparisons remain the primary quantitative basis for method evaluation.

5. Conclusions

This study developed and evaluated a reproducible WRF-to-WBGT post-processing framework using three component pathways, four correction branches, three WRF grid spacings, two independent observation sources, and 70 consecutive summer validation days in central Texas. WRF supplied the meteorological forcing; WBGT was calculated afterward from T_n , T_g , and T_a . For that reason, the product should be described as WRF-derived WBGT rather than as a native WRF output.

Liljegren provided the strongest general full-day framework. Against KestrelMet/AWN, d01 Liljegren Carter_FAO had the lowest aggregate RMSE and the highest bootstrap winner frequency. However, its paired RMSE advantage over d01 Hunter Carter_FAO was not resolved at the 95% level. Against Tempest, d01 Liljegren FAO and d01 Liljegren Carter were statistically indistinguishable co-leaders. These results support a physically based Liljegren core while also showing that one correction branch should not be presented as universally best.

Hunter/Minyard performance depended strongly on the T_n treatment. Hunter Base substantially overestimated T_n and WBGT in this dataset. After the Carter/Bernard substitution, Hunter became one of the leading configurations. Hunter Carter_FAO was especially competitive for daily maxima.

Dimiceli/Piltz remained useful as a simplified Tg pathway, but its generally cool globe-temperature behavior limited its performance as part of a complete WBGT estimate.

The FAO wind-height adjustment reduced the large WRF wind bias but did not reproduce the strongly sheltered observed wind regime by itself. WRF also showed a dry dew-point bias and positive shortwave-radiation bias. These errors act in different directions in the WBGT calculation. Due to this, acceptable final WBGT results may therefore occur from compensation among imperfect meteorological inputs. Component and driver diagnostics are necessary if the model-derived WBGT is going to be interpreted physically.

Horizontal resolution did not produce a steady improvement from d01 to d03. The 25-km domain performed best for the strongest full-day configurations, while the 1-km domain became competitive for daily maxima. Grid spacing should therefore be selected with the target application and exposure in mind; the finest WRF nest should not automatically be assumed to provide the best point WBGT.

Observation-source uncertainty was also large enough to affect the final ranking. KestrelMet/AWN and Tempest differed by about 2.27°F on common WBGT timestamps, which is far larger than several of the paired differences among the leading model configurations. The most defensible recommendation is therefore source-aware and multi-metric: use Liljegren as the preferred general WRF-derived WBGT framework; retain Hunter Carter_FAO as a strong alternative, particularly for peak applications; evaluate FAO and Carter branches in the context of local exposure; and avoid declaring one method superior when the paired whole-day bootstrap interval includes zero. The framework is reproducible and is ready for follow-on work on sheltered microclimate representativeness and independent open-field validation.

Figure Captions

Fig. 1 Validation site and instrument exposure in San Marcos, Texas. (A) Aerial view of the residential property and surrounding neighborhood. (B-H) Ground-level views of the KestrelMet/AWN and Tempest stations installed over maintained grass within the backyard. The photographs illustrate nearby fencing, mature trees, and the adjacent residence, all of which contribute to the partially sheltered exposure discussed in Section 2.2

Fig. 2 WRF-derived WBGT processing workflow used in Paper 1. WRF meteorological fields from d01, d02, and d03 were sampled at the grid point nearest the validation site, converted to core driver variables, routed through psychrometric wet-bulb and wind treatments, passed through the Hunter/Minyard, Dimiceli/Piltz, and Liljegren component pathways, combined with Base, FAO, Carter, and Carter_FAO branches, and then compared with observations using project-level statistics and bootstrap ranking

Fig. 3 Top 10 full-day WBGT methods by RMSE against KestrelMet/AWN observations. Error bars denote the 2.5th to 97.5th percentile bootstrap RMSE range, and right-side labels show the bootstrap winner frequency

Fig. 4 Top 10 full-day WBGT methods by RMSE against Tempest observations. Error bars denote the 2.5th to 97.5th percentile bootstrap RMSE range, and right-side labels show the bootstrap winner frequency

Fig. 5 Best full-day WBGT RMSE by method family for KestrelMet/AWN and Tempest validation sources. Lower values indicate better agreement with the corresponding observation source

Fig. 6 Best full-day WBGT RMSE by correction branch for KestrelMet/AWN and Tempest validation sources. The branch comparison highlights the impact of the Base, FAO, Carter, and Carter_FAO treatments on final WBGT performance

Fig. 7 Best full-day WBGT RMSE attained within each domain for KestrelMet/AWN and Tempest validation sources. For each domain, the plotted value represents the lowest full-day RMSE found among all pathway and branch combinations

Fig. 8 Best natural wet-bulb RMSE attained within each domain using KestrelMet/AWN observations. These values summarize how well the different post-processing pathways reproduced the observed natural wet-bulb component

Fig. 9 Best black-globe-temperature RMSE attained within each domain using KestrelMet/AWN observations. Lower RMSE indicates better agreement with the observed globe-temperature component

Fig. 10 Meteorological-driver RMSE by domain for validation against KestrelMet/AWN observations. Each panel uses the native units of the indicated variable so that air temperature, dew point, wind, solar radiation, and station pressure can be interpreted without mixing unlike units on one axis

Fig. 11 Meteorological-driver RMSE by domain for validation against Tempest observations. Each panel uses the native units of the indicated variable; the panel structure matches Figure 10 for direct source-to-source comparison

Fig. 12 Top 10 daily-peak WBGT methods by RMSE against KestrelMet/AWN observations. Error bars denote the 2.5th to 97.5th percentile bootstrap RMSE range, and right-side labels show the bootstrap winner frequency

Fig. 13 Top 10 daily-peak WBGT methods by RMSE against Tempest observations. Error bars denote the 2.5th to 97.5th percentile bootstrap RMSE range, and right-side labels show the bootstrap winner frequency

Fig. 14 Best daily-peak WBGT RMSE attained within each domain for KestrelMet/AWN and Tempest validation sources. For each domain, the plotted value represents the lowest daily-peak RMSE found among all pathway and branch combinations

Fig. 15 Hourly WBGT RMSE from 0700 to 1900 local time for the best full-day method in each domain, using KestrelMet/AWN observations. The plot shows how error evolved through the diurnal validation window

Fig. 16 Hourly WBGT RMSE from 0700 to 1900 local time for the best full-day method in each domain, using Tempest observations. The plot shows how error evolved through the diurnal validation window

Appendix A. Reproducibility Details for the WBGT Calculation

A1. Solar partition and atmospheric emissivity

The Hunter/Kuehn and Dimiceli/Piltz globe calculations used a fixed direct/diffuse shortwave partition of 0.67/0.33. Atmospheric thermal emissivity followed $\epsilon_a = 0.575 e_a^{(1/7)}$, where e_a is vapor pressure in hPa. The Hunter and Dimiceli/Piltz globe calculations used the same basic atmospheric-emissivity treatment so that differences in T_g could be traced primarily to the globe formulation rather than to an unrelated emissivity choice.

A2. Hunter/Kuehn and Dimiceli/Piltz globe calculations

The Hunter/Kuehn globe temperature was obtained by solving a nonlinear radiative-convective energy balance in which absorbed shortwave and atmospheric longwave radiation are balanced by thermal emission and wind-dependent convection. The Dimiceli/Piltz implementation used the published tangent-line approximation to the globe solution. Both calculations used the pathway wind selected for the branch; therefore, Base/Carter used WRF 10-m wind, while FAO/Carter_FAO used the FAO-adjusted wind.

A3. Liljegren iterative solution

The Liljegren wick calculation iteratively updated air properties, saturation vapor pressure, latent heat, net radiation, convective heat transfer, and mass transfer until the temperature solution converged. The Liljegren globe calculation separately iterated the radiative and convective balance for the black globe. Dynamic solar partitioning was recalculated from shortwave radiation, solar zenith angle, and Earth-Sun distance at each valid time. This explicit treatment is one reason Liljegren was retained as the primary physically based benchmark in the study.

A4. Output nomenclature

Each branch exported WBGT_<family>_<branch>_F, Tn_<family>_<branch>_F, and Tg_<family>_<branch>_F. Family labels were Hunter, DimiceliPiltz, and Liljegren; branch labels were Base, FAO, Carter, and Carter_FAO. The naming convention was used consistently in the daily files, project statistics, bootstrap outputs, figures, and manuscript tables so that every reported result could be traced back to one calculation pathway and branch.

Appendix B. Suggested Supplementary Material

To keep the main manuscript focused, the exhaustive 36-candidate metric tables, daily statistics, full pairwise bootstrap outputs, and project-level quality-control summaries should remain in supplementary material. Recommended supplements include complete full-day metrics for both observation sources, daily-maximum magnitude and timing tables, complete natural-wet-bulb and black-globe component statistics, meteorological-driver metrics, the full bootstrap-ranking tables, and project-level quality-control summaries.

References

- Allen RG, Pereira LS, Raes D, Smith M (1998) Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements. FAO Irrigation and Drainage Paper 56. Food and Agriculture Organization of the United Nations, Rome
- Budd GM (2008) Wet-bulb globe temperature (WBGT) - its history and its limitations. *Journal of Science and Medicine in Sport* 11(1):20-32. <https://doi.org/10.1016/j.jsams.2007.07.003>
- Carter AW, Zaitchik BF, Gohlke JM, Wang S, Richardson MB (2020) Methods for estimating wet bulb globe temperature from remote and low-cost data: A comparative study in Central Alabama. *GeoHealth* 4(5):e2019GH000231. <https://doi.org/10.1029/2019GH000231>
- Clark J, Konrad CE (2020) Accuracy Assessment of Experimental Wet Bulb Globe Temperature Forecasts Across North Carolina and the Continental United States. *Carolinas Integrated Sciences and Assessments / NOAA Southeast Regional Climate Center*, 11 pp
- Clark J, Konrad CE (2024) Observations and estimates of wet-bulb globe temperature in varied microclimates. *Journal of Applied Meteorology and Climatology* 63(2):305-319. <https://doi.org/10.1175/JAMC-D-23-0078.1>
- Dimiceli VE, Piltz SF, Amburn SA (2011) Estimation of black globe temperature for calculation of the wet bulb globe temperature index. *Proceedings of the World Congress on Engineering and Computer Science 2011, Vol. II*, San Francisco, California, 591-599

905 Grundstein A, Williams C, Phan M, Cooper E (2015) Regional heat safety thresholds for athletics in the
 906 contiguous United States. *Applied Geography* 56:55-60. <https://doi.org/10.1016/j.apgeog.2014.10.014>
 907 Hunter CH, Minyard CO (1999) Estimating Wet Bulb Globe Temperature Using Standard Meteorological
 908 Measurements. WSRC-MS-99-00757, Westinghouse Savannah River Company, Aiken, South
 909 Carolina
 910 Iacono MJ, Delamere JS, Mlawer EJ, Shephard MW, Clough SA, Collins WD (2008) Radiative forcing
 911 by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *Journal of*
 912 *Geophysical Research: Atmospheres* 113:D13103. <https://doi.org/10.1029/2008JD009944>
 913 Kain JS (2004) The Kain-Fritsch convective parameterization: An update. *Journal of Applied*
 914 *Meteorology* 43:170-181. [https://doi.org/10.1175/1520-0450\(2004\)043<0170:TKCPAU>2.0.CO;2](https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2)
 915 Liljegren JC, Carhart RA, Lawday P, Tschopp S, Sharp R (2008) Modeling the wet bulb globe
 916 temperature using standard meteorological measurements. *Journal of Occupational and*
 917 *Environmental Hygiene* 5(10):645-655. <https://doi.org/10.1080/15459620802310770>
 918 Nakanishi M, Niino H (2009) Development of an improved turbulence closure model for the atmospheric
 919 boundary layer. *Journal of the Meteorological Society of Japan* 87(5):895-912.
 920 <https://doi.org/10.2151/jmsj.87.895>
 921 Niu G-Y, Yang Z-L, Mitchell KE, Chen F, Ek MB, Barlage M, Longuevergne L, Manning K, Niyogi D,
 922 Tewari M et al (2011) The community Noah land surface model with multiparameterization options
 923 (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of*
 924 *Geophysical Research: Atmospheres* 116:D12109. <https://doi.org/10.1029/2010JD015139>
 925 Nielsen-Kellerman (2025) KestrelMet 6400 WBGT Wi-Fi Weather Station Instruction Manual.
 926 Document 319115_1_25.10.29
 927 Skamarock WC, Klemp JB, Dudhia J, Gill DO, Liu Z, Berner J, Wang W, Powers JG, Duda MG, Barker
 928 DM, Huang X-Y (2019) A Description of the Advanced Research WRF Version 4. NCAR Technical
 929 Note NCAR/TN-556+STR, 145 pp. <https://doi.org/10.5065/1dfh-6p97>
 930 Thompson G, Field PR, Rasmussen RM, Hall WD (2008) Explicit forecasts of winter precipitation using
 931 an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization.
 932 *Monthly Weather Review* 136(12):5095-5115. <https://doi.org/10.1175/2008MWR2387.1>
 933 WeatherFlow-Tempest (2026) Tempest Weather System Technical Specifications. WeatherFlow-
 934 Tempest, tempest.earth

935 **Statements and Declarations**

936 **Competing Interests**

937 None

938 **Author Contributions**

939 William Hatheway: Conceptualization, methodology, software, validation, formal analysis, investigation,
940 data curation, visualization, and writing

941 **Data and Code Availability**

942 The analysis workflow uses Python and shell scripts to extract WRF point data, calculate the 12 WBGT
943 family/branch combinations, retrieve and time-match observations, merge daily datasets, calculate
944 project-level statistics, and run the whole-day bootstrap.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryDataScripts.zip](#)