Effect of WBGT Index Measurement Location on Heat Stress Category Classification

SAMUEL N. CHEUVRONT1, ELIZABETH M. CARUSO1, KRISTEN R. HEAVENS1, ANTHONY J. KARIS1, WILLIAM R. SANTEE1, CHRIS TROYANOS2, and PIERRE D’HEMECOURT2

1U.S. Army Research Institute of Environmental Medicine, Natick, MA; and 2Boston Athletic Association, Boston, MA

ABSTRACT

CHEUVRONT, S. N., E. M. CARUSO, K. R. HEAVENS, A. J. KARIS, W. R. SANTEE, C. TROYANOS, and P. D’HEMECOURT. Effect of WBGT Index Measurement Location on Heat Stress Category Classification. Med. Sci. Sports Exerc., Vol. 47, No. 9, pp. 1958–1964, 2015. The location of the wet bulb globe temperature (WBGT) index measurement may affect heat stress flag category classification. **Purpose:** This study aimed to compare WBGT measurements at three locations along the Boston Marathon race course and compare WBGT estimates for meteorological stations and 72-h advanced WBGT forecasts. **Methods:** WBGT was measured hourly from 1000 to 1400 h at approximately 7 km, approximately 18 km, and approximately 30 km on the Boston Marathon race course. Simultaneous WBGT estimates were made for two meteorological stations southeast of the course via a commercial online system, which also provided 72-h advanced forecasts. **Results:** The measurement difference (mean ± SD) among course locations was 0.2°C ± 1.8°C WBGT (ANOVA, *P* > 0.05). The difference between course and stations was 1.9°C ± 2.4°C WBGT (*t*-test, *P* < 0.05). Station values underestimated (*n* = 98) or overestimated (*n* = 13) course values by >3°C WBGT (>0.5 flag category) in 111 of 245 paired comparisons (45%). Higher black globe and lower wet bulb temperatures explained over- and underestimates, respectively. Significant underestimates of WBGT resulted in misclassification of green (labeled white) and black (labeled red) course flag categories (χ², *P* < 0.05). Forecast data significantly underestimated red (labeled amber) and black (labeled red) course flag categories. **Conclusions:** Differences in WBGT index along 23 km of the Boston Marathon race route can be small enough to warrant single measurements. However, significant misclassification of flag categories occurred using WBGT estimates for meteorological stations; thus, local measurements are preferred. If the relation between station WBGT forecasts and the race sites can be established, the forecast WBGT values could be corrected to give advanced warning of approximate flag conditions. Similar work is proposed for other venues to improve heat stress monitoring. **Key Words:** HEAT STRAIN, EXERCISE, WEATHER, MARATHON RUNNING

The wet bulb globe temperature (WBGT) index is an environmental heat stress measure developed in the 1950s (26) and has been adopted for use by military (25), occupational (Occupational Safety and Health Administration) (19), sports medicine communities (1,12,22), and others to provide guidance for mitigating heat strain when conducting training or performing exercise. It is also an important measurement for estimating the effect of environment on exercise performance (10,17). The equation for calculating the outdoor WBGT index is 0.7 *T*<sub>nwb</sub> + 0.2 *T*<sub>bg</sub> + 0.1 *T*<sub>db</sub> (26), where *T*<sub>nwb</sub> is a natural wet bulb thermometer exposed to sun and wind, *T*<sub>bg</sub> is a black globe thermometer similarly exposed, and *T*<sub>db</sub> is a shaded dry bulb thermometer. Thus, the relative influence of each measurement is weighted 70%, 20%, and 10%, respectively. The *T*<sub>db</sub> is essentially equivalent to the shaded air temperature (*T*<sub>a</sub>) when measured at 1.2 m above ground. All three components of the WBGT are nonstandard weather instruments, and the *T*<sub>nwb</sub> additionally requires maintenance. Recommendations for care and use of the non-standard WBGT components can be found elsewhere (25), but modern technologies are making WBGT measurements increasingly simple.

Several small, portable, and user-friendly all-in-one devices for monitoring the WBGT index have been developed to simplify its use. The extent to which portable devices have been validated varies (3,15,24). An alternative to portable systems includes estimating the WBGT index from models that calculate the required WBGT inputs from standard meteorological station measurements of *T*<sub>a</sub>, humidity, wind speed, barometric pressure, and, when available, radiant heat load (3,13,15,20,24). One notable benefit of models that use standard meteorological measurements is that WBGT can also be potentially forecast. One limitation of models is the inability of commercial users to easily implement their use.
Another is the significant geographical distance between meteorological station(s) and the location(s) of interest for monitoring.

WBGT index values are traditionally measured at a single location to inform the user community about potential heat stress hazards; thus, it is recommended that the WBGT be measured as close as possible to the planned activity to capture the actual exposure conditions that will affect health (1,12,25). When the monitoring area is large (e.g., a 42-km road race), it is recommended that WBGT be monitored at multiple locations to account for spatial variations (15). This is often not possible. The degree to which location can produce errors large enough to affect flag category (risk) classification has only been minimally quantified and exclusively so with military WBGT heat stress categories in mind (15,20,24). The sports medicine WBGT heat stress categories have broader bands (5°C–8°C; mean, 6°C) (1,12) than those in military medicine (1.1°C–1.7°C) (25), and they are shifted downward relative to military doctrine, presumably to account for higher anticipated rates of metabolic heat production (i.e., >600 W) (1,12). Like military and occupational situations, sporting events often take place in venues affording shade or shelter from wind (e.g., stadiums, buildings, fields) (11) and may be subject to the influences of sport-specific surface phenomena (blacktop or artificial turf surfaces) (4). Local microclimates may therefore differ significantly from the prevailing standard conditions for meteorological observations that provide full, uniform sensor exposures (8). Presently, there are little data examining the effect that location might have on WBGT heat stress category classification, particularly in sports medicine.

The purpose of this study was to compare locally measured WBGT index values using newly available portable WBGT monitors (Kestrel 4400; Nielsen-Kellerman, Boothwyn, PA) at three locations along the historic Boston Marathon race route. We also sought to compare the WBGT at each location along the route with modeled (and forecast) WBGT values calculated from meteorological station data nearest each course location using an online commercial system (Schneider Electric USA, Inc., Omaha, NE). The effect of location on WBGT heat stress category classification (1,12) was a primary interest.

**RESEARCH DESIGN AND METHODS**

**Preliminary procedures.** A series of pilot studies was conducted to evaluate the performance of the field-expedient WBGT index measurement device used in this study (Kestrel 4400; Nielsen-Kellerman, Boothwyn, PA). First, sports medicine WBGT flag categories of white (<10°C), green (10°C–18°C), amber (18°C–23°C), red (23°C–28°C), and black (28°C) (1,12) were simulated indoors using an environmental chamber. The desired flag category conditions and limited comparisons with the WBGT measured on the course, a larger-scale study was proposed to test the efficacy of the system alongside “ground truth” comparisons made at different locations along the course.

**Experimental procedures.** The Boston Marathon is run every year on Patriot’s Day at a time of day when the

**WBGT INDEX LOCATION AND HEAT STRESS**

Medical & Science in Sports & Exercise, 1959
course is exposed to a solar zenith angle >45° for most of the race. For this reason, data collection for this study occurred during the meteorological summer (between the spring and fall equinox). All measurements were made hourly from 1000 to 1400 h to approximate the time of the day and exposure duration for most race participants. The three Kestrel devices were positioned on an anthropometric scale (23) 1.2 vertical meters above the ground (dirt) and 1.2 horizontal meters from the road (asphalt) to approximate the environmental stress experienced by runners at chest height on the course. A rotating tripod vane was used for hands-free monitoring and constant orientation of sensors with the prevailing wind. Instrument stabilization was standardized by allowing a 30-min equilibration period before the first outdoor measurement. The Kestrel devices were placed at 7 km on the north side of the road (Ashland), 18 km on the south side of the road (Natick), and 30 km on the north side of the road (Newton). The locations for the WBGT index measurements along the course (km) were selected by the BAA. The specific positions for the Kestrels were determined by necessity on the basis of race day restrictions, but the chosen positions were maintained throughout the study. There were no differences in roadside surfaces, and shade was not a factor at the test sites, given the testing season and times of the day.

Coordinates for the three course locations were used as settings for Schneider Electric monitoring. Weather data for the Ashland and Natick sites were pulled from the nearest airport weather station in Norwood (24 km southeast of Ashland and 18 km southeast of Natick; ground elevation, 15 m above sea level). Weather data for Newton came from the nearest observatory weather station in East Milton (15 km southeast of Newton; ground elevation, 200 m above sea level). All thermometer and hygrometer readings were taken at 1.5 m, and anemometer readings were taken at 10 m above each respective ground elevation. The distances among towns on the course and from points of interest on the course to the nearest meteorological stations are similar to the distances between WBGT monitors and meteorological stations modeled in other studies (15) and similar to the distances between many military WBGT stations and actual posttraining sites.

All WBGT index input measurements were made in real time at the top of every hour. Kestrel data were documented electronically at 5-min intervals (not 5-min averages) for later trend analysis. However, the 5-min value at the top of every hour was used for comparisons to match the hourly Schneider Electric sampling rate, which was accessed in real time using a smartphone and documented by hand. Forecast data were retrieved similarly 72 h in advance for 1000, 1200, and 1400 h. The hourly WBGT components for Schneider Electric had to be derived because of limited online system outputs (\(T_a\) and relative humidity). To estimate the relative agreement between course and meteorological station WBGT components, \(T_a\) was considered equivalent to \(T_{db}\) and the psychrometric wet bulb (\(T_{wwb}\)) was calculated from relative humidity, \(T_a\), and barometric pressure using a psychrometric formula (2). The online WBGT broadcast value was then used to solve for the unknown \(T_{wb}\), which was calculated indirectly by Schneider Electric (9) but similarly not displayed as output.

**Statistical analysis.** WBGT index comparisons were first made by conventional inferential analysis. Mean differences among course locations were compared by one-way repeated measures ANOVA (Tukey post hoc), and mean differences between course and meteorological stations were compared by an independent t-test. Individual differences >3°C WBGT (>0.5 flag category) were considered meaningful (a priori) and were tabulated and expressed as a percentage of the whole to quantify the quality of agreement. The practical application of the differences between course and meteorological station conditions was examined using the chi-square goodness-of-fit test (21) to determine whether the observed flag category frequencies (meteorological stations) differed significantly from the expected ground truth flag categories (course). Chi-square post hoc tests were performed in accordance with the methods of Markowski and Markowski (16). Finally, linear regression and bootstrap regression analyses were used to examine how well the forecast WBGT predicted the course WBGT. The number of observations required to detect meaningful (>3°C) differences in WBGT (ANOVA and t-test) and approximately 25% differences in the relative frequencies of flag categories (chi-square) were both met at the 0.05 alpha and 0.80 power levels (5). Data were analyzed using the GraphPad Prism software (version 5.04; GraphPad, San Diego, CA) except where otherwise indicated and expressed as mean ± SD.

**RESULTS**

Table 1 compares WBGT index measurements using the three Kestrel weather stations against an environmental chamber set to white, green, amber, red, and black flag heat stress categories. The mean difference between any Kestrel and the environmental chamber sensors was ≤0.7°C WBGT index. The mean difference among Kestrel instruments positioned within 0.6 m of each other was ≤0.2°C WBGT. The difference among the three Kestrels positioned similarly outdoors on a hot day (WBGT, 26°C) was ≤0.5°C WBGT. The slightly larger differences outdoors may have been due to differences in the distances between WBGT monitors and meteorological stations across the range of WBGT flag categories

<table>
<thead>
<tr>
<th>Flag Category</th>
<th>Ashland Kestrel</th>
<th>Natick Kestrel</th>
<th>Newton Kestrel</th>
<th>Chamber</th>
</tr>
</thead>
<tbody>
<tr>
<td>White (&lt;10)</td>
<td>10.6</td>
<td>10.7</td>
<td>10.6</td>
<td>10.3</td>
</tr>
<tr>
<td>Green (10–18)</td>
<td>17.9</td>
<td>18.1</td>
<td>17.9</td>
<td>17.4</td>
</tr>
<tr>
<td>Amber (18–23)</td>
<td>22.2</td>
<td>22.4</td>
<td>22.2</td>
<td>22.1</td>
</tr>
<tr>
<td>Red (23–28)</td>
<td>26.8</td>
<td>27.0</td>
<td>26.8</td>
<td>26.8</td>
</tr>
<tr>
<td>Black (&gt;28)</td>
<td>31.4</td>
<td>31.5</td>
<td>31.4</td>
<td>31.5</td>
</tr>
</tbody>
</table>

Data represent mean values measured over 30 min. Measurement details are provided in text.
to variable wind versus fixed air velocity (indoors) or radiant load. The results of Kestrel chamber tests at 28°C WBGT before any field measurements were similar and stable throughout the study. The individual components of the WBGT ($T_{nwb}$, $T_{bg}$, and $T_{db}$) were nearly identical among Kestrels, and their mean was very similar when compared with that in the chamber measures (see Table, Supplemental Digital Content, Comparison of WBGT components using Kestrel devices to environmental chamber settings, http://links.lww.com/MSS/A485).

Hourly course data were collected ($n = 5$) at each location ($n = 3$) for 17 d for a total of 255 potential data pairs. The mean WBGT index differences among locations were $0.2°C \pm 1.8°C (P > 0.05)$. Figure 1A–C displays 85 ($5 \times 17 = 85$) individual town differences in WBGT in each of three panels (A–C). Only 11 (4.3%) of 255 individual differences were $>3°C$ WBGT or approximately 0.5 flag category. The 11 errant data pairs were explained primarily by $8°C$–$14°C$ differences in $T_{bg}$ owing to transient variations in cloud cover. $T_{db}$ and $T_{nwb}$ were extremely uniform among course locations (difference, $\leq3°C$). As a result of the 95.7% agreement, the three locations were collapsed into one large group. Meteorological station measures among towns were also not different ($0.4°C \pm 1.6°C; P > 0.05$) and were therefore collapsed similarly.

Ten hourly course versus meteorological station data pairs were lost because of satellite signal problems. The mean WBGT index difference between the 245 remaining data pairs was $1.9°C \pm 2.4°C$ WBGT ($P < 0.05$). Meteorological station measures underestimated ($n = 98$) or overestimated ($n = 13$) course measures by $>3°C$ WBGT in 111 of 245 paired comparisons (45.3%). WBGT sensor inputs were measured at three time points only (1000, 1200, and 1400 h) for a total of 153 paired comparisons. Higher meteorological station $T_{bg}$ and lower $T_{nwb}$ explained over- and underestimates, respectively (Fig. 2). There were 20 white, 50 green, 66 amber, 62 red, and 47 black flag heat stress categories among the 245 data pairs. Significant underestimates of the WBGT resulted in misclassification of green (labeled white) and black (labeled red) flag categories ($\chi^2, P < 0.05$), but differences never exceeded one flag category. Figure 3 represents a plot of the relation between course and meteorological station WBGT measures. Flag category misclassification can be visualized graphically by data pairs that fall outside the shaded regions on the grid.

The WBGT index forecast was obtained 72 h in advance for hourly ($n = 3$) course measurements made at each location ($n = 3$) for 15 d for a total of 135 potential paired comparisons. Six data pairs were lost because of satellite signal problems. Of the 129 remaining data pairs, 15 were white, 15 were green, 41 were amber, 39 were red, and 19 were black flag heat stress categories. The 72-h forecast WBGT data were not significantly different from day of meteorological station WBGT estimates, but they significantly underestimated ($n = 36$) or overestimated ($n = 16$) course measures by $>3°C$ WBGT in 54 of 129 paired comparisons

FIGURE 1—Mean differences in WBGT among course locations along the Boston Marathon race route in Ashland (approximately 7 km), Natick (approximately 18 km), and Newton (approximately 30 km) were not significant ($P > 0.05$). Individual differences within the shaded band ($3°C$ WBGT), or approximately 0.5 flag category, were also considered marginal. Data represent 85 paired comparisons per panel.
model shrinkage (6) and error (18) were considered acceptable and stable. The formulated equation was therefore considered sufficiently accurate to forecast WBGT within 0.5 flag category at the Boston Marathon. A small-scale validation of the equation was performed by applying it to the Natick town forecast (Norwood Airport) obtained 72 h before the 2014 Boston Marathon. “Ground truth” measures were made in Natick with a Kestrel device. The original forecast consistently underestimated the WBGT measured on race day by $1^\circ$C (WBGT), but the correction $T_{\text{see}}$ captured the measured value at four of five time points, underestimating one time point (1300) by $2^\circ$C WBGT (Fig. 4).

**DISCUSSION**

The purposes of this study were to compare the WBGT index measured simultaneously at three different locations along the Boston Marathon race route and also to compare...
each location with real-time measures and forecasts made at meteorological stations to determine how location affects WBGT heat stress category classification. This study evaluated extensively the performance of the Kestrel 4400 WBGT monitor and compared the Kestrel data with the Schneider Electric online WBGT monitoring service, both of which recently became available for commercial use in sports, military, and wilderness medicine settings.

One major finding of this study was that the measured WBGT index did not differ significantly (P > 0.05) or meaningfully (only 4.3% of values >3°C) at 7, 18, and 30 km along the Boston Marathon race route (Fig. 1A–C). Strong evidence for the validity of WBGT measures on the course (Table 1 and Results) lends essential support to this conclusion. It is important to acknowledge that comparisons like these are highly dependent on the locations being compared. The middle half of the Boston Marathon race route seems to be relatively dependent on the locations being compared. The middle half of the Boston Marathon race route (Fig. 1A–C). Strong evidence for the validity of WBGT measures on the course (Table 1 and Results) lends essential support to this conclusion. It is important to acknowledge that comparisons like these are highly dependent on the locations being compared. The middle half of the Boston Marathon race route seems to be relatively dependent on the locations being compared.

A second major finding of this study was that when remote meteorological station (airport or observatory) data were used to model WBGT index, the differences were significant (P < 0.05) and meaningful compared with course WBGT index measures (45.3% of values >3°C). The practical outcome borne of this difference was a significant underestimate of green and black flag conditions by one flag category each (Fig. 3). In general, the underestimate occurred as a result of lower meteorological station \( T_{wb} \) (Fig. 2B), which is the most heavily weighted WBGT equation parameter (70%). The errors might have been larger were it not for the frequently much higher meteorological station \( T_{bg} \) (Fig. 2C). Differences in sensor heights between meteorological observations (1.5–10 m above ground) and those measured on an anthropometric scale (1.2 m above ground) may have contributed to the observed differences in both \( T_{wb} \) and \( T_{bg} \) (8,23). Differences in wind at anthropometric (1.2 m) and nonanthropometric scales (10 m) would, in particular, affect both measures. Indeed, even the position of the meteorological sensors above sea level (15–199 m) might have contributed, given that all three sensors on the course were uniform at approximately 70 m above sea level. Similar observations have been made by others comparing ground with tower measurements (15). The Schneider Electric indirect calculation of the \( T_{bg} \) (9) may have played a role, and our derivation of the \( T_{pwb} \) (from relative humidity and \( T_d \)) on the assumptions that \( T_d = T_{db} \) from Schneider Electric display outputs may have introduced error. However, the course-with-meteorological station WBGT comparison was without assumptions and the differences between course- and meteorological station-derived WBGT in this study were larger than those in other studies where the distance between measurements was 6–9 km (15). Therefore, differences in geography, season, or some other factor(s) may contribute to differences in model error. On the basis of these findings, it is recommended that the WBGT be measured on location when practical. WBGT index values predicted from regional weather service data (models) should be used cautiously when portable devices are not available.

A third important finding of this study was that the 72-h Schneider Electric WBGT index forecast performed similarly to Schneider Electric day of meteorological station WBGT measures, thus affording reasonable WBGT predictions (within one flag category) for use with sports medicine flag categories. Caution is warranted, however, because predictions based on the forecasts significantly underestimated the more severe red and black flag categories. The source of error in the forecast is likely multivariate (\( T_{wb} \), \( T_{bg} \), other). In the specific case of the Boston Marathon, an equation describing and incorporating the bias (1 SEE) effectively shrinks the forecast error to within 0.5 flag category of the course value. The small remaining errors observed during the improved forecast validation at the 2014 Boston Marathon (Fig. 4), where WBGT was still slightly underestimated, was likely due to the unavoidable microclimate created by the crowd of spectators and the presence of so many runners near the Kestrel WBGT monitoring device on the race day.

One limitation to the interpretations of this study is the WBGT index flag category used. In sports medicine, each WBGT heat stress category spans ≥6°C (1), whereas those used by the military span only 1.1°C–1.7°C (25). It is likely that greater misclassifications would have resulted from smaller category spans (20). Although the practical importance of 1°C–2°C WBGT category differences may seem small and probably makes no difference within the context of the sports medicine flag categories, they can have a significant effect in very hot and humid environments, especially when protective clothing is worn (23) or when work/rest cycles are required (25). Conditions above a sports medicine red flag (>28°C) are grounds for canceling athletic events (1,12), but the same WBGT is only a green flag in the military, where less intense activity (<600 W) is expected to continue as long as the appropriate work/rest cycles are used (25). Another limitation was the absence of a true natural wet bulb temperature (thermometer covered with wet wick). Chamber, Kestrel, and meteorological station \( T_{wb} \) (or \( T_{pwb} \)) were ultimately calculated from relative humidity. Although low wind velocities may be problematic, waterless \( T_{wb} \) measures have previously been shown to agree well with wet wick measures (24). Modeled differences between indirect and direct \( T_{wb} \) measures, even over extreme humidity ranges (10%–90%), also produce WBGT differences smaller than 0.5 sports medicine flag category (7). Although direct findings from this study may be applied conceptually without restriction, they should be understood specifically within the context of a sports
medicine application and the location, instruments, and models used herein. Importantly, the results of this study nevertheless provide useful, practical, and insightful information for approaching environmental heat stress monitoring during exercise or work.

CONCLUSIONS

This study suggests that differences in WBGT index along 23 km of the Boston Marathon race route can be small enough to warrant single measurements. However, significant indelity in WBGT modeled from proximate meteorological stations can occur and misclassification of heat stress can be off by a full category. On the basis of the relation between meteorological station and race site WBGT values, the modeling can be adjusted to predict/forecast sports medicine heat flag categories for the Boston Marathon to within 0.5 flag category, but the specific adjustments may not be applicable to other locations nor adequate for the more narrowly defined military or occupational flag classifications of heat illness dangers. Direct local measurements of WBGT are therefore preferred when possible. Similar work is proposed for other venues to improve heat stress monitoring.

We would like to thank Stephen P. Mullen, Bruce S. Cadarette, Timothy J. Driscoll, Jeffrey D. O’Mara, and Robert W. Kenefick for their critical technical assistance with this study. We also recognize the technical support afforded to us by Dale Nelson (Schneider Electric), Tommy Loomis and Michael McNaughton (Kestrel), and especially the BAA.

This work was funded by the United States Army Medical Research and Materiel Command.

The opinions or assertions contained herein are the private views of the authors and should not be construed as official or reflecting the views of the Army or the Department of Defense. Any citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement of approval of the products or services of these organizations. This article is approved for public release; distribution is unlimited. The authors have no conflicts of interest to report.

The results of this study do not constitute endorsement by the American College of Sports Medicine.

REFERENCES