Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

PAPER

OPEN ACCESS

RECEIVED
24 April 2025

REVISED
22 July 2025

ACCEPTED FOR PUBLICATION
1 October 2025

PUBLISHED
14 October 2025

Core body temperature estimation from heart rate via multi-model Kalman filtering and variance-based fusion

Yuanzhe Zhao[1] and **Jeroen HM Bergmann**[1,2,*]

[1] Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[2] Department of Technology & Innovation, University of Southern Denmark, Odense, Denmark
[*] Author to whom any correspondence should be addressed.

E-mail: jberg@iti.sdu.dk

Keywords: core temperature, body temperature, Kalman filter, core temperature estimation, heart rate

Abstract

Objective. Accurate and non-invasive estimation of core body temperature (CBT) is essential for preventing heat-related illnesses during physical activity and thermal stress. The objective of this work is to develop and evaluate a framework for real-time CBT estimation using only heart rate (HR) data, enabling a lightweight solution suitable for deployment on wearable devices. *Approach.* We propose a multi-model Kalman filtering (KF) framework with variance-based fusion. Two variants were developed: a supervised Physiological State-Specific KF (PSSK) that uses activity labels (rest, exercise, recovery) to train distinct models, and an unsupervised trial clustering-based KF (TCBK) that clusters trials based on HR–CBT features to capture latent physiological variability without state annotations. Both models were evaluated on two independent datasets and compared against baseline methods. *Main results.* In within-dataset evaluations, TCBK achieved the highest accuracy with a root mean square error (RMSE) of 0.38 °C (Dataset 1) and 0.41 °C (Dataset 2). In cross-dataset generalization, PSSK demonstrated superior robustness with an RMSE of 0.88 °C, whereas the TCBK model's error increased to 1.56 °C. Both proposed models outperformed the established Buller and Falcone models. *Significance.* This work demonstrates that lightweight, HR-only models can provide accurate CBT estimation by incorporating state- or context-aware modeling. The framework offers a practical and deployable solution for continuous thermal strain monitoring in occupational and athletic settings, providing a balance between performance and real-world applicability for wearable technology.

1. Introduction

Core body temperature (CBT) is a vital physiological indicator of thermal strain, particularly during physical activity or exposure to heat (Venugopal *et al* 2016, Zhang *et al* 2025). In high-risk environments such as firefighting, military operations, and endurance sports, elevated CBT is strongly associated with heat-related illnesses including heat exhaustion, heat stroke, and exertional heat illness (Jardine 2007, Gauer and Meyers 2019). These conditions can develop rapidly and unpredictably, often in the absence of early subjective symptoms. If left unrecognized, dangerously high CBT levels may result in impaired cognitive and physical performance, loss of consciousness, organ failure, or even death (Arbury *et al* 2014, Hess *et al* 2014, Faurie *et al* 2022). Consequently, continuous and accurate CBT monitoring is critical not only for safeguarding individual health, but also for enabling data-driven strategies to manage thermal risk, guide exertion levels, and prevent life-threatening outcomes in heat-stressed environments (Laxminarayan *et al* 2014, Garzön-Villalba *et al* 2017, Buller *et al* 2021).

Traditional methods of CBT assessment—such as rectal or esophageal thermometry, and ingestible telemetric pills—offer high accuracy but are invasive, cumbersome, and impractical for routine or prolonged use in the field (Moran and Mendal 2002, Soehle *et al* 2020, Dolson *et al* 2022, Tokizawa *et al* 2022). More accessible alternatives like axillary (underarm) and oral temperature measurements have been shown to poorly correlate with true CBT during periods of physical exertion or thermal challenge.

These methods tend to underestimate actual thermal strain, making them unreliable for monitoring dynamic changes in core temperature (CT) (Hooper and Andrews 2006, Casa *et al* 2007, Ganio *et al* 2009). As a result, there is a pressing need to develop and deploy effective, non-invasive CBT monitoring technologies that are suitable for real-world applications in occupational, athletic, and military settings.

Over the years, two broad approaches have emerged to address the challenge of non-invasive CBT estimation. One relies on physiological modeling grounded in the biophysics of human thermal regulation. Classical thermoregulatory models such as those by Stolwijk (1971) and Fiala (1999) simulate human thermal responses by dividing the body into multiple anatomical segments and tissue layers. These models compute dynamic heat transfer via conduction, convection, radiation, evaporation, and metabolic heat production, integrating inputs such as environmental conditions, metabolic rate, skin blood flow, and sweat rate (Fu *et al* 2016, Yang *et al* 2017, Welles *et al* 2018a). While physiologically detailed and useful for predictive simulation, these models require extensive parameterization, including clothing insulation and individual anthropometry, and are computationally intensive. Their assumptions of average physiological responses limit adaptability to individual variability, which is crucial for real-time, personalized monitoring in uncontrolled environments (Xu and Werner 1997, Yokota *et al* 2012).

The second approach leverages wearable sensing technologies to infer CBT indirectly through measurable physiological signals. Wearable systems offer the potential for continuous, unobtrusive, and real-time CBT estimation, making them highly attractive for field applications (Moyen *et al* 2021, Dolson *et al* 2022, Kubota *et al* 2025, Zhao and Bergmann 2025). Among various biosignals, heart rate (HR) has gained particular attention as a surrogate input for CBT due to its strong association with thermoregulatory activitys (Wyss *et al* 1974, Rubin 1987, Horn *et al* 2013).

The physiological relationship between HR and CBT reflects both thermoregulatory and metabolic demands. Under passive heat stress, rising CBT triggers thermoregulatory responses-such as increased skin blood flow and sweating-requiring higher cardiac output and thus elevating HR. Furthermore, during physical exercise, HR increases rapidly to meet metabolic demand, and the resulting internal heat production leads to a progressive rise in CBT (Rowell 1974, Karvonen and Vuorimaa 1988). This empirical positive correlation between HR and CBT has been leveraged in various CBT estimation models, some of which rely solely on HR, while others incorporate additional physiological signals, using techniques such as linear regression (Niedermann *et al* 2014, Richmond *et al* 2015), machine learning and neural networks (Verdel *et al* 2021, Han *et al* 2025), and state-space approaches like Kalman filters (KFs) (Welles *et al* 2018b, Rizvi 2022).

Among these, KFs are particularly well-suited for real-time CBT tracking due to their ability to integrate prior knowledge of system dynamics with noisy observations. A typical KF model consists of a state transition function, which models the temporal evolution of CBT, and an observation function, which relates the observed HR to the hidden CBT state (Buller *et al* 2013, Khodarahmi and Maihami 2023). The recursive structure of the KF enables real-time updates as new measurements arrive, making it attractive for wearable implementation.

Buller *et al* (2013, 2015) proposed one of the earliest KF-based approaches for CBT estimation using HR. Their method uses an extended KF (EKF) with a fixed quadratic observation model and a first-order autoregressive state model, assuming uniform physiological dynamics across individuals and activity phases. More recently, Falcone *et al* (2024) introduced the Biphasic KF-Based (BKFB) model, which modifies Buller's approach by dividing the estimation into two distinct phases-heating and cooling-based on predefined HR thresholds. Each phase uses a separate model to account for differences in thermoregulatory behavior during CBT increase and decrease.

While both methods enable real-time, non-invasive CBT estimation, they rely on two key assumptions: (1) a universal HR-CBT mapping that ignores individual variability, and (2) clearly separable physiological states identifiable through simple HR thresholds. However, in real-world conditions, different states (e.g. rest and recovery) may yield similar HR patterns, and individuals exhibit substantial variability in thermal response. These limitations reduce model flexibility and generalizability.

To address these issues, we propose a novel multi-model KF framework with variance-based fusion for CBT estimation. In our framework, multiple independently parameterized KFs are executed in parallel, and their outputs are dynamically fused based on posterior variance to produce a single, robust CBT estimate. We introduce two complementary implementations of this framework:

(i) Physiological state-specific KF (PSSK) leverage activity labels (e.g. *Rest*, *Exercise*, *Recovery*) to train separate models for each physiological state. This method assumes that activity annotations are available or can be reliably inferred.
(ii) Trial clustering-based KFs (TCBK) operate in a fully unsupervised manner by treating each participant-condition trial as an independent unit. A set of summary physiological features is

extracted from each trial, including the mean, minimum, and maximum values of HR and CBT, as well as the slope of their linear relationship. K-means clustering is applied to group trials by latent physiological patterns, and cluster-specific KFs are trained accordingly.

Both variants share a common fusion strategy but differ in how model specialization is achieved. While PSSK benefits from label-aware modeling when activity segmentation is available, TCBK provides a label-free alternative that can adapt to both individual and contextual variability. We evaluate both models on two independent datasets and compare their performance against baseline approaches under within-dataset and cross-dataset conditions.

2. Method

2.1. Datasets

This study employed datasets from two previously studies involving healthy adult participants exposed to controlled thermal and physical activity conditions.

Dataset 1 includes 18 participants (10 males and 8 females) aged between 19 and 36 years (Richmond *et al* 2015, Havenith *et al* 2024). All sessions were conducted indoors. Each participant completed two experimental sessions corresponding to different clothing conditions: one with permeable cotton coverall and one with an impermeable coated nylon coverall. In each session, participants followed a structured protocol that began with 10 min of seated rest, followed by 40 min of treadmill walking. Afterward, participants rested until their rectal temperature had decreased by at least 0.4°C from the peak value. The session ended with the second 40 min treadmill walking period. Throughout the experiment, HR, rectal temperature (as the ground truth for CT), and activity codes indicating the participant's physiological state (e.g. rest or exercise) were continuously recorded.

Dataset 2 includes 13 male participants aged between 18 and 45 years (Eggenberger *et al* 2018). Each participant completed two experimental sessions (heat stages 1 and 2), designed to simulate thermal strain under different workload and clothing conditions. In both stages, participants first rested seated outdoors for 15 min, then rested indoors for an additional 15 min. This was followed by a cycling session lasting between 20 and 60 min indoors. The exercise was terminated if any of the following conditions were met: rectal temperature increased by at least 1.5°C from baseline, rectal temperature exceeded 38.5°C, voluntary fatigue was reached, or a maximum of 75 min had elapsed. Finally, participants rested seated outdoors for 30 min. In heat stage 1, participants exercised at 75% of their maximal HR ($\mathrm{HR}_{\mathrm{max}}$) wearing light athletic clothing (T-shirt and shorts). In heat stage 2, the intensity was reduced to 50% of $\mathrm{HR}_{\mathrm{max}}$, and participants wore full firefighter protective gear. HR and rectal temperature were continuously recorded throughout each stage. No activity labels were provided in this dataset.

The demographic characteristics of participants in both datasets including average age, body mass, height, and body fat percentage are summarized in table 1, along with details of the environmental temperature and humidity conditions, activity structure, and physiological variables used in this study.

2.2. Multi-state KFs

To estimate CBT non-invasively from HR, we design a multi-model estimation framework built upon customized KFs and a robust fusion strategy. This section presents two complementary implementations: one based on predefined physiological states, and another relying on unsupervised participant-level clustering. Both approaches share a common structure: multiple independently parameterized KFs (Buller *et al* 2013, Khodarahmi and Maihami 2023) run in parallel, and their outputs are fused using a variance-based weighting mechanism to produce a final temperature estimate.

Each model k consists of a state transition model and a observation model:

State transition model:

$$T_t^{(k)} = a_1^{(k)} T_{t-1}^{(k)} + a_0^{(k)} + \epsilon_t^{(k)}, \quad \epsilon_t^{(k)} \sim \mathcal{N}\left(0, \gamma_k^2\right). \tag{1}$$

The state transition model (equation (1)) is formulated as a first-order autoregressive process. This reflects the physiological principle of thermal inertia, where the body's CT changes gradually, and its state at any given time is highly dependent on its immediate past. In this equation, $T_t^{(k)}$ denotes the estimated CT at time t under model k, and $T_{t-1}^{(k)}$ is the previous estimate. The coefficients $a_1^{(k)}$ and $a_0^{(k)}$

Table 1. Summary of datasets. M = male, F = Female and HR = Heart rate.

Attribute	Dataset 1 (Richmond *et al* 2015, Havenith *et al* 2024)	Dataset 2 (Eggenberger *et al* 2018)
Number of participants	18 (10 M/8 F)	13 (13 M)
Age range (years)	19–36	18–45
Mean age (±SD) in years	25.1(±5.5)	30.9(±5.4)
Mean body mass (±SD) in kg	69.7(±11.5)	77.5(±6.1)
Mean height (±SD) in cm	174.9(±6.5)	179.2(±6.4)
Mean body fat (±SD) in%	14.9(±5.2)	13.1(±4.3)
Activity protocol & Environmental temperature and humidity	10 min rest(40 °C, 20%), 40 min walk(40 °C, 20%), rest until $\Delta T \geqslant$ 0.4 °C (22 °C, 50%), 40 min walk(40 °C, 20%)	15 min rest(23 °C, 25%), 15 min rest(35 °C, 57%), 20–60 min cycling(35 °C, 57%), 30 min rest(23 °C, 25%)
Stage conditions	permeable clothing(Stage 1), impermeable clothing(Stage 2)	75% of $\mathrm{HR_{max}}$, light clothing (Stage 1), 50% of $\mathrm{HR_{max}}$, firefighter gear (Stage 2)
Used variables	HR, rectal temperature, activity code	HR, rectal temperature

are learned from data and capture the autoregressive dynamics of temperature evolution. The process noise $\epsilon_t^{(k)}$ is assumed to follow a zero-mean Gaussian distribution with variance γ_k^2.

Observation model:

$$\hat{\mathrm{HR}}_t^{(k)} = b_2^{(k)}\left(T_t^{(k)}\right)^2 + b_1^{(k)}T_t^{(k)} + b_0^{(k)} + \eta_t^{(k)}, \quad \eta_t^{(k)} \sim \mathcal{N}\left(0, \sigma_k^2\right). \tag{2}$$

The observation model (equation (2)) is based on the principle, established by Buller *et al*, that HR can serve as a powerful, albeit noisy, observation of the CT. The physiological justification is that HR is a leading indicator of thermal strain, as it encapsulates critical information about both metabolic heat production and heat transfer (Buller *et al* 2013). Therefore, our model leverages this readily available signal to infer the hidden state of CT at each time step. In this equation, $\hat{\mathrm{HR}}_t^{(k)}$ is the predicted HR based on the estimated temperature $T_t^{(k)}$ under model k. Coefficients $b_2^{(k)}$, $b_1^{(k)}$, and $b_0^{(k)}$ are learned separately for each model k. The observation noise $\eta_t^{(k)}$ is assumed Gaussian with zero mean and variance σ_k^2.

At the beginning of the estimation process, each KF is initialized with a CBT of 37.1° C and an initial variance of 0.01 °C^2. Subsequently, each filter updates its CT estimate and associated uncertainty through a recursive process composed of prediction and update steps:

1. Prediction step

$$\hat{T}_t^{(k)} = a_1^{(k)}T_{t-1}^{(k)} + a_0^{(k)} \tag{3}$$

$$\hat{v}_t^{(k)} = v_{t-1}^{(k)} + \gamma_k^2. \tag{4}$$

In this step, the prior estimate of the CT at time t under model k, denoted by $\hat{T}_t^{(k)}$, is predicted from the previous posterior estimate $T_{t-1}^{(k)}$ using the linear state transition model. The associated predicted variance $\hat{v}_t^{(k)}$ represents the uncertainty of the prior estimate, obtained by propagating the previous variance $v_{t-1}^{(k)}$ and adding the process noise variance γ_k^2.

2. Observation prediction

$$\hat{\mathrm{HR}}_t^{(k)} = b_2^{(k)}\left(\hat{T}_t^{(k)}\right)^2 + b_1^{(k)}\hat{T}_t^{(k)} + b_0^{(k)}. \tag{5}$$

The expected HR $\hat{\mathrm{HR}}_t^{(k)}$ is computed from the prior temperature estimate $\hat{T}_t^{(k)}$ using the nonlinear observation model. The quadratic form reflects the physiological relationship between HR and CT under model k (Buller *et al* 2013).

3. EKF mapping function variance coefficient

$$m_t^{(k)} = \frac{\mathrm{d}\hat{\mathrm{HR}}}{\mathrm{d}T} = 2b_2^{(k)}\hat{T}_t^{(k)} + b_1^{(k)}. \tag{6}$$

To enable Kalman-based updates with a nonlinear observation model, the function is linearized and the term $m_t^{(k)}$ denotes the EKF mapping function variance coefficient.

4. Kalman gain computation

$$k_t^{(k)} = \frac{\hat{v}_t^{(k)} \cdot m_t^{(k)}}{\left(m_t^{(k)}\right)^2 \cdot \hat{v}_t^{(k)} + \sigma_k^2}. \tag{7}$$

The Kalman gain $k_t^{(k)}$ is a crucial parameter that is dynamically computed at each time step. Its primary role is to determine the weight given to the new observation (the HR measurement) versus the model's prediction when updating the state estimate.It is derived from the predicted state variance $\hat{v}_t^{(k)}$, the EKF mapping function variance coefficient $m_t^{(k)}$, and the observation noise variance σ_k^2. Specifically, a larger Kalman gain reflects higher confidence in the observation and results in a stronger correction to the state estimate. Conversely, a smaller gain implies greater trust in the model's prediction, leading to a more conservative update. By continuously adjusting based on evolving uncertainty in both the model and measurements, the Kalman gain plays a central role in balancing model dynamics with real-time sensor data.

5. State update

$$T_t^{(k)} = \hat{T}_t^{(k)} + k_t^{(k)} \cdot \left(\mathrm{HR}_t - \hat{\mathrm{HR}}_t^{(k)}\right). \tag{8}$$

The posterior estimate of CT $T_t^{(k)}$ is obtained by correcting the prior estimate $\hat{T}_t^{(k)}$ with the innovation term $\mathrm{HR}_t - \hat{\mathrm{HR}}_t^{(k)}$, which reflects the discrepancy between the actual and predicted HR. The correction is scaled by the Kalman gain.

6. Variance update

$$v_t^{(k)} = \left(1 - k_t^{(k)} \cdot m_t^{(k)}\right) \cdot \hat{v}_t^{(k)}. \tag{9}$$

The posterior variance $v_t^{(k)}$ is updated to reflect the reduced uncertainty after incorporating the current observation. The reduction factor $1 - k_t^{(k)} \cdot m_t^{(k)}$ indicates the information contribution of the observation relative to the prior uncertainty.

After each KF corresponding to model k independently completes the prediction and update steps, we obtain multiple parallel estimates $T_t^{(k)}$, each associated with a posterior variance $v_t^{(k)}$. To produce a unified and robust estimate of CBT, we fuse these outputs using a variance-based weighting strategy that reflects the confidence of each model.

$$w_k = \frac{1/v_t^{(k)}}{\sum_j 1/v_t^{(j)}} \tag{10}$$

$$T_t^{\mathrm{fused}} = \sum_k w_k \cdot T_t^{(k)}. \tag{11}$$

Here, w_k is the normalized fusion weight for model k, computed as the inverse of its posterior variance. This inverse-variance weighting scheme gives higher importance to models with greater confidence (i.e. lower uncertainty). The final fused estimate T_t^{fused} is thus a weighted average of all model-specific outputs, representing the best overall prediction under the assumption of independent Gaussian uncertainties.

2.2.1. PSSKs

The dynamic relationship between HR and CBT varies depending on the underlying physiological state. Leveraging the activity labels available in Dataset 1, we categorize the data into three distinct states: *rest*, *exercise*, and *recovery*. For each physiological state $k \in \{\mathrm{rest}, \mathrm{exercise}, \mathrm{recovery}\}$, we independently estimate a state transition model and a nonlinear observation model, as described in section 2.2.

At inference time, three independent KFs-each corresponding to a specific physiological state-are executed in parallel. All filters receive the same HR input but apply their own state-specific dynamics and observation models. Each filter outputs an updated estimate of CT along with its associated posterior variance. To obtain the final prediction, the outputs of all filters are fused using an inverse-variance weighting strategy, which reflects the relative confidence of each model.

2.2.2. TCBK filters

To account for both individual variability and condition-dependent differences in thermoregulatory response, we propose a data-driven framework called **TCBK Filters**. Unlike traditional participant-level clustering, TCBK treats each *trial*-defined as one participant completing a session under a specific condition-as an independent unit for clustering and model training.

For each trial, we compute a set of summary features from the corresponding time series, including the mean, minimum, and maximum values of HR and CT, as well as the slope of their linear relationship. These features are standardized via z-score normalization, and K-means clustering is applied to group trials into $K = 3$ clusters based on physiological similarity. To ensure optimal clustering quality, we perform an exhaustive search over all possible combinations of two or more features and select the subset that minimizes the downstream CT estimation root mean squared error (RMSE).

Once the trial clusters are formed, we train a separate KF for each cluster. Each model shares the same structure as described in section 2.2, with cluster-specific parameters for the state transition and observation models: $\{a_1^{(c)}, a_0^{(c)}, \gamma_c, b_2^{(c)}, b_1^{(c)}, b_0^{(c)}, \sigma_c\}$. These parameters are estimated using linear regression and second-order polynomial regression.

During inference, all cluster-specific KFs operate in parallel, regardless of the original cluster assignment of the test trial. Each filter produces an independent estimate of CBT and its associated uncertainty. The final temperature estimate is then obtained via inverse-variance weighted fusion, where models with lower uncertainty contribute more heavily to the prediction.

This trial-level clustering approach provides greater flexibility and adaptability than participant-based clustering, as it captures variations across both individuals and experimental conditions. It is particularly advantageous in datasets where activity labels are unavailable or unreliable, and where physiological response patterns are highly condition-specific.

2.3. Model evaluation methods

To quantitatively assess the accuracy of CT estimation, this study employed three metrics: mean absolute error (MAE), RMSE, and Bland–Altman statistics. These metrics were computed by comparing estimated CT values against the ground truth at each time point (Giavarina 2015, Zhao and Bergmann 2023).

All metrics were calculated based on concatenated time-series data from all participants within each dataset, providing a comprehensive evaluation of group-level performance. This approach reflects the overall estimation accuracy across diverse individuals and conditions.

Additionally, Bland–Altman plots were used to visualize and analyze prediction residuals, offering insights into both systematic biases (mean difference (MD)) and the spread of errors (standard deviation (SD)). These plots also revealed temperature-dependent error patterns, making them particularly valuable for assessing model behavior across the full physiological range, especially at elevated CTs.

MAE:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{T}_i - T_i\right|. \tag{12}$$

RMSE:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{T}_i - T_i\right)^2}. \tag{13}$$

MD and SD:

$$\mathrm{MD} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{T}_i - T_i\right) \tag{14}$$

$$\mathrm{SD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(\hat{T}_i - T_i\right) - \mathrm{MD}\right]^2}. \tag{15}$$

These statistics are used in Bland–Altman analysis to evaluate the agreement between the estimated and reference temperatures. MD reflects systematic bias, while SD indicates the spread of the residual errors.

All metrics are reported based on concatenated predictions across all participants in the same dataset, enabling group-level performance evaluation.

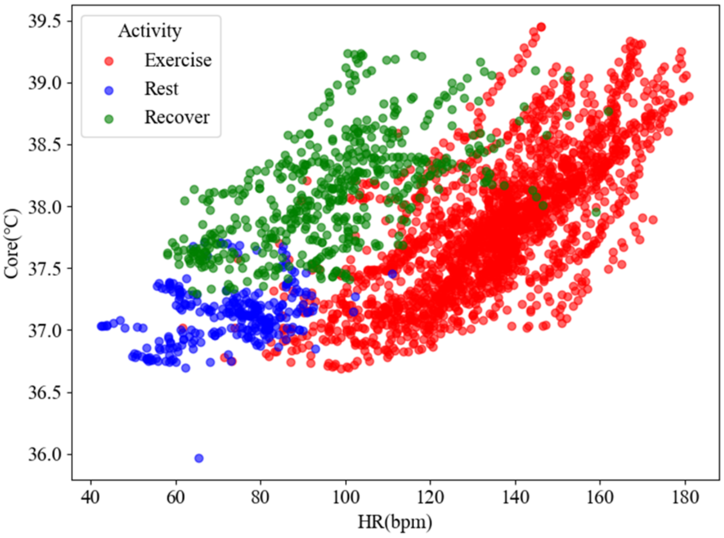


Figure 1. Heart rate, core temperature, and activity states in Dataset 1.

Table 2. Model parameters for PSSK, TCBK, Buller, and Falcone methods on Dataset 1.

Model	State/cluster	a_1	a_0	γ	b_2	b_1	b_0	σ
PSSK	Rest	1.0203	−0.7684	0.04 732	−13.1008	988.3444	−18 562.3877	11.6849
	Exercise	1.0049	−0.1602	0.04 166	−6.2478	499.4499	−9816.3406	13.4916
	Recovery	1.0027	−0.1179	0.03 080	−9.8319	776.9867	−15 232.9121	15.3268
TCBK	Best feature set	*Max_Core, Min_Core*						
	Cluster 1	1.0038	−0.1273	0.0322	−8.9938	709.4774	−13 844.8504	23.6935
	Cluster 2	0.9957	0.1727	0.0470	2.3500	−144.4650	2227.7564	24.7168
	Cluster 3	0.9663	1.2802	0.1119	−31.6569	2393.5826	−45 118.5013	22.5549
Buller	—	1.0033	−0.1114	0.0447	−8.2635	652.4964	−12 733.1383	24.2347
Falcone	HR threshold	*120 bpm*						
	Rising phase	0.9983	0.0928	0.0410	0.0142	13.7887	−401.3093	11.1216
	Falling phase	0.9799	0.7481	0.0430	0.8260	−57.6493	1094.6583	16.9798

3. Results

3.1. Evaluation on dataset 1

Figure 1 presents an overview of Dataset 1, including HR, CT, and their associated activity states (rest, recovery, exercise). The figure reveals that the physiological signals corresponding to different activity states exhibit overlap, indicating the difficulty of state discrimination based solely on HR and CT.

Table 2 summarizes the model parameters used in the four approaches evaluated on Dataset 1. For the PSSK model, separate KF parameters were trained for each physiological state. For TCBK, the optimal feature subset used for participant clustering is listed, along with the parameters for each resulting cluster-specific model. Buller's model was fitted using a global parameter set, while the Falcone model includes separate parameter sets for the rising and falling phases, along with the best-performing HR threshold used for phase switching.

Figure 2 illustrates the estimated CBT over time using four different models on Dataset 1. The predictions from PSSK, TCBK, Buller's model, and Falcone's biphasic Kalman model are plotted alongside the ground truth values.

Table 3 compares the performance of all four methods on Dataset 1 using RMSE, MAE, MD, and SD metrics. Both PSSK and TCBK outperform the baseline methods in all aspects.

Among the proposed methods, TCBK achieves the lowest RMSE (0.38 °C), MAE (0.29 °C), and SD (0.38 °C), indicating its superior accuracy and consistency. PSSK ranks second in RMSE (0.42 °C), MAE (0.32 °C), and SD (0.42 °C), but achieves the best MD (0.01 °C).

In contrast, Buller's model shows moderate estimation accuracy, but with a negative bias (MD = –0.06 °C) and higher variance (SD = 0.43 °C). Falcone's model performs the worst overall, with the

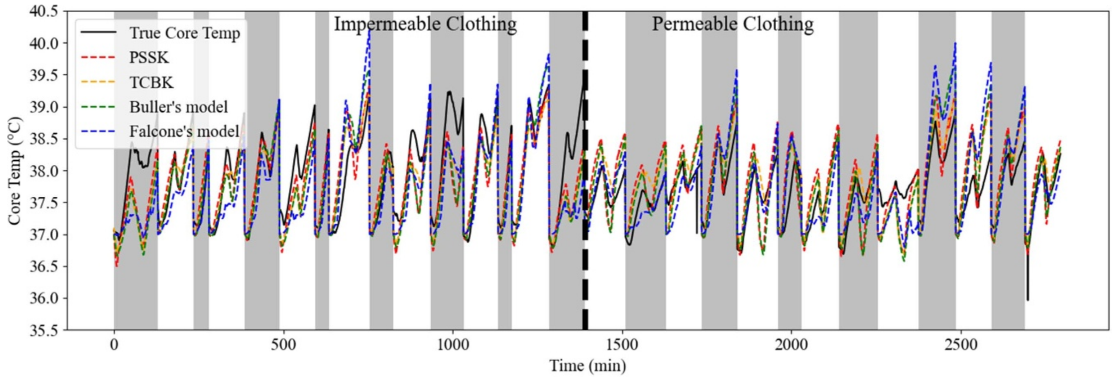
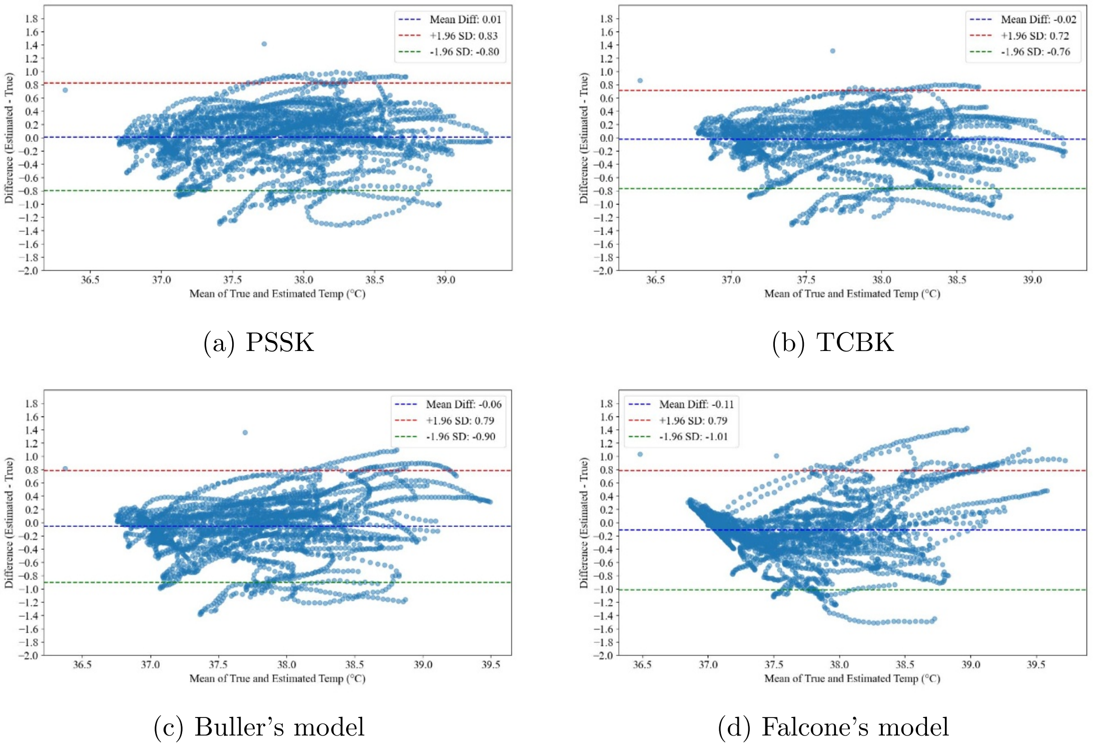


Figure 2. Prediction results on Dataset 1 using different models: The figure presents time-series plots of estimated core body temperature alongside the ground truth across all participants under two clothing conditions. Alternating gray and white background regions distinguish individual participants, enabling visual comparison of model performance across subjects. A vertical dashed line separates the two clothing conditions-impermeable on the left and permeable on the right.

Table 3. Performance comparison of CBT estimation methods on Dataset 1. The lowest obtained values are shown in bold.

Method	RMSE (°C)	MAE (°C)	MD (°C)	SD (°C)
PSSK	0.42	0.32	**0.01**	0.42
TCBK	**0.38**	**0.29**	–0.02	**0.38**
Buller *et al*	0.43	0.33	–0.06	0.43
Falcone *et al*	0.47	0.35	–0.11	0.46



(a) PSSK

(b) TCBK

(c) Buller's model

(d) Falcone's model

Figure 3. Bland–Altman plots of core temperature estimation on Dataset 1 using different models. The plots show the mean difference and 95% limits of agreement between predicted and ground truth temperatures.

highest RMSE (0.47 °C) and the largest negative bias (MD $= -0.11$ °C), reflecting underestimation of CT.

Figure 3 presents the Bland–Altman plots for the four models, showing the agreement between estimated and ground truth CTs. PSSK and TCBK show tighter agreement and narrower limits of agreement compared to Buller's and Falcone's methods, especially at higher temperatures.

Table 4. Model parameters for TCBK, Buller, and Falcone methods on Dataset 2.

Model	State/Cluster	a_1	a_0	γ	b_2	b_1	b_0	σ
TCBK	Best feature set				*Avg_Core,Avg_HR, Min_Core,Max_HR*			
	Cluster 1	1.0007	−0.0231	0.0150	−34.4348	2594.7056	−48 746.2651	34.6130
	Cluster 2	0.9993	0.0267	0.0162	−11.5513	880.5294	−16 664.2499	27.1740
	Cluster 3	0.9987	0.0495	0.0153	−23.0177	1759.9794	−33 519.5462	28.8622
Buller	—	0.9998	0.0086	0.0156	−10.6568	816.6215	−15 521.8354	22.3432
Falcone	HR threshold					*100 bpm*		
	Rising phase	1.0000	0.0044	0.0154	−22.5097	1692.6964	−31 678.6174	17.8573
	Falling phase	0.9969	0.1145	0.0151	−0.5451	55.5554	−1234.3091	9.4995

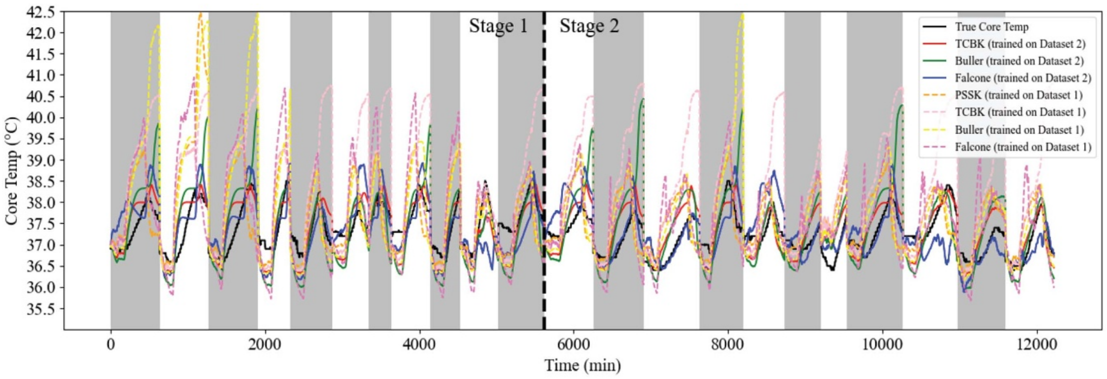


Figure 4. Prediction results on Dataset 2 using different models. The figure shows estimated and true core temperatures for all participants under different workload and clothing conditions.

3.2. Evaluation on dataset 2

This section presents the evaluation results on Dataset 2, which consists exclusively of male participants under two thermal workload stages with distinct clothing conditions. Unlike Dataset 1, no activity labels were available, and the physiological states were not explicitly segmented. Thus, the PSSK model trained on Dataset 1 was directly transferred to Dataset 2 without retraining. For TCBK, Buller's, and Falcone's methods, new models were trained using Dataset 2. In total, six methods were evaluated:

- Participant clustering-based KFs (TCBK, trained on Dataset 2)
- Buller's model (trained on Dataset 2)
- Falcone's model (trained on Dataset 2)
- PSSK (trained on Dataset 1)
- TCBK (trained on Dataset 1)
- Buller's model (trained on Dataset 1)
- Falcone's model (trained on Dataset 1)

Table 4 reports the model parameters for the methods trained on Dataset 2. The parameters for the transferred models are shown in table 2 from section 3.1.

Figure 4 illustrates the estimated CBT across all participants in Dataset 2 using the six methods. Alternating gray and white background regions denote different participants, allowing visual comparison of model performance across individuals. The time-series plot demonstrates how well each model tracks the true temperature trend, highlighting differences in generalization performance for models transferred from Dataset 1 versus those trained on Dataset 2.

Table 5 compares the estimation accuracy of all seven models on Dataset 2 using RMSE, MAE, MD, and SD of residuals. Among the models trained directly on Dataset 2, the TCBK method achieves the best overall performance, with the lowest RMSE (0.41 °C), MAE (0.35 °C), and SD (0.41 °C). Falcone's model, while slightly inferior in RMSE and MAE, exhibits the lowest MD (0.01 °C), indicating minimal systematic bias. Buller's model performs worse than both TCBK and Falcone in all metrics, with a higher residual spread (SD = 0.68 °C) and moderate bias (MD = 0.10 °C).

In contrast, all models trained on Dataset 1 show significant performance degradation when applied to Dataset 2. The transferred TCBK model suffers the largest errors, with RMSE = 1.56 °C and MAE

Table 5. Performance comparison of CBT estimation methods on Dataset 2. The lowest values are shown in bold.

Method	RMSE (°C)	MAE (°C)	MD (°C)	SD (°C)
TCBK (trained on Dataset 2)	**0.41**	**0.35**	0.06	**0.41**
Buller (trained on Dataset 2)	0.69	0.52	0.10	0.68
Falcone (trained on Dataset 2)	0.50	0.39	**0.01**	0.50
PSSK (trained on Dataset 1)	0.88	0.68	0.36	0.80
TCBK (trained on Dataset 1)	1.56	1.25	1.09	1.12
Buller (trained on Dataset 1)	1.13	0.79	0.43	1.04
Falcone (trained on Dataset 1)	0.97	0.74	0.33	0.91

$= 1.25\,^\circ$C, suggesting poor generalization of the learned clusters to unseen conditions. Buller's and Falcone's models trained on Dataset 1 also show elevated RMSEs (1.13 °C and 0.97 °C respectively), though slightly better than TCBK. PSSK performs the best among transferred models, achieving an RMSE of 0.88 °C and MAE of 0.68 °C, but still with large bias (MD = 0.36 °C) and variance (SD = 0.80 °C). This indicates that while PSSK retains some robustness, its performance still falls short compared to models retrained on Dataset 2.

Figure 5 presents the Bland–Altman plots for all seven models. Among the models trained on Dataset 2, the TCBK method demonstrates the narrowest limits of agreement and the least variability across the full temperature range, indicating greater agreement across the range. Falcone's model also shows relatively stable performance, though with slightly larger dispersion. In contrast, Buller's model exhibits wider scatter and a more pronounced positive bias, particularly at higher CTs.

For the models transferred from Dataset 1, all exhibit larger limits of agreement and increased error variability. The transferred TCBK model shows the highest bias and variance, reflecting poor generalization under unseen thermal conditions. PSSK maintains closer alignment to the ground truth in terms of error trend, though it tends to overestimate CT at certain time points. Buller's transferred model produces substantial overestimations in the high-temperature range. Although Falcone's transferred model shows more uniformly distributed residuals, its overall error magnitude remains considerable.

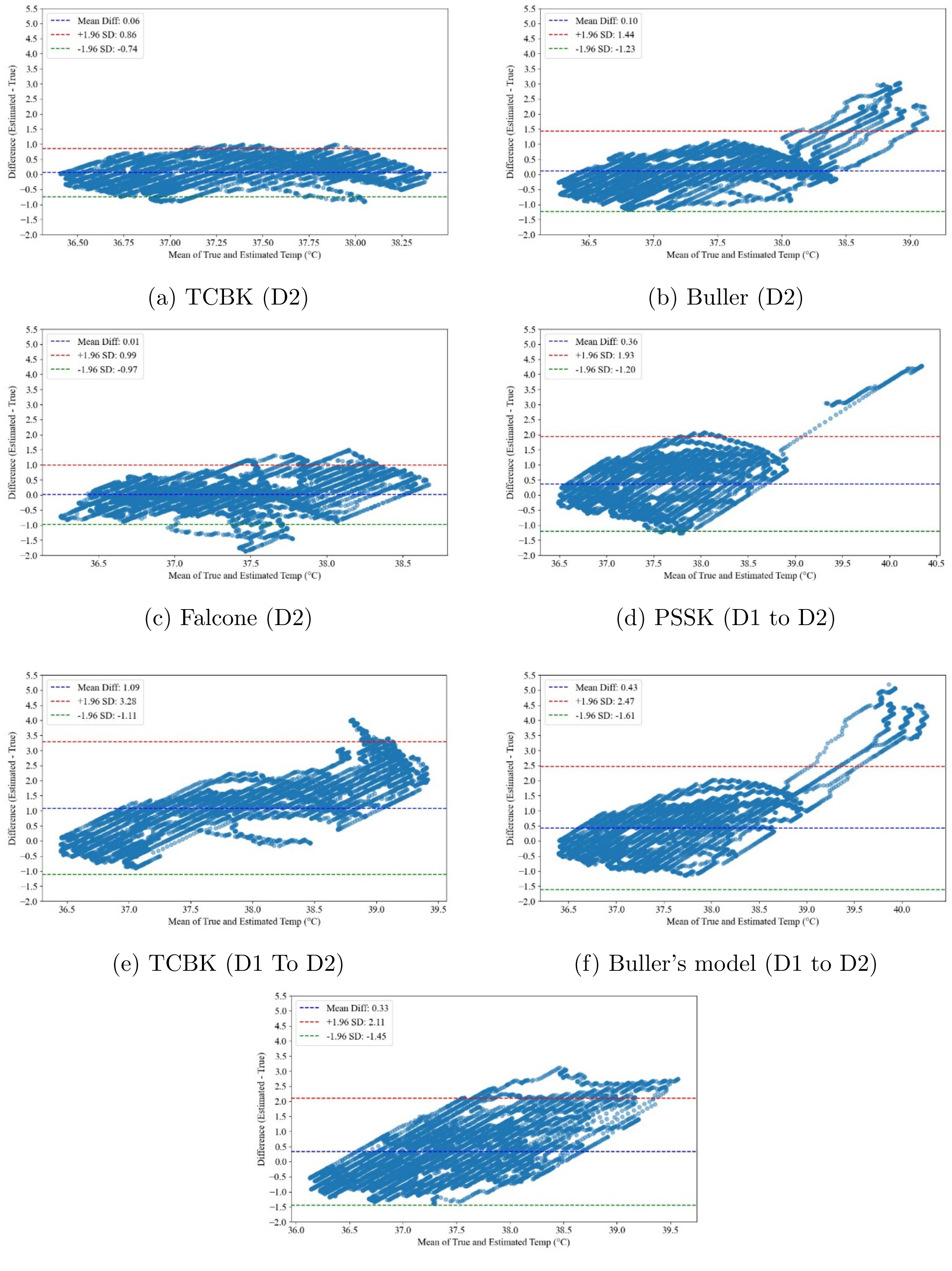
4. Discussion

We proposed two models for CBT estimation: the PSSK and the TCBK filter. Both models are designed for implementation on wearable sensor devices. The PSSK model is trained using HR, CT, and physiological state labels (e.g. rest, exercise, recovery), while the TCBK model is trained based on clustered trials that capture participant-specific and environmental variability.

During real-time CBT estimation, both models rely solely on HR as input variable, which offers several advantages. Numerous commercially available wearable devices now provide reliable HR monitoring (Fuller *et al* 2020). Moreover, relying exclusively on HR minimizes computational burden and sensor requirements, making the system lightweight and suitable for deployment on low-power wearable platforms. Compared to multimodal approaches that incorporate skin temperature or heat flux, HR-only models are easy to implement and any additional sensors can interfere with comfort and usability. It is also worth noting that skin temperature can vary considerably depending on the sensor location and often deviates substantially from CT (Casa *et al* 2007, Ganio *et al* 2009, Zhao *et al* 2024), which can reduce the consistency and generalizability of multimodal models.

However, as shown in figure 1, the relationship between HR and CBT varies across different physiological states. The method proposed by Buller *et al* assumes a static, one-to-one mapping between HR and CBT, which overlooks such variability and leads to substantial estimation error even within a single dataset. Falcone *et al* attempted to address this by introducing phase-specific modeling, dividing the CBT trajectory into up-phase and down-phase based on HR thresholds. However, figure 1 also demonstrates that there is considerable overlap in the HR–CBT relationship across different physiological states, indicating that such binary division is insufficient to capture the complexity of the underlying dynamics. As a result, Falcone's model fails to generalize effectively and, in Dataset 1, even performs worse than Buller's simpler model.

In contrast, our PSSK model explicitly captures state-dependent dynamics by training distinct KFs for each physiological state and combining their predictions using confidence-weighted fusion. This approach accommodates both the diversity across states and similarities in overlapping regions. Consequently, PSSK consistently outperforms both baseline models in within-dataset and cross-dataset evaluations, demonstrating its capacity for better physiological representation.



(a) TCBK (D2)

(b) Buller (D2)

(c) Falcone (D2)

(d) PSSK (D1 to D2)

(e) TCBK (D1 To D2)

(f) Buller's model (D1 to D2)

(g) Falcone (D1 to D2)

Figure 5. Bland–Altman plots of core temperature estimation on Dataset 2 using different models. The plots show the agreement between predicted and ground truth values for models trained on Dataset 1 and Dataset 2.

To address the challenges of requiring real-time state labels, we developed the TCBK approach, which eliminates the need for explicit activity state annotations by clustering trials using descriptive features derived from HR and CBT signals. These include the mean, minimum, maximum, and slope of their linear relationship. This allows TCBK to model inter-individual and contextual variability, such as differences in baseline fitness or environmental stressors, without relying on direct physiological state labels. This contributes to TCBK's strong performance in within-dataset evaluations.

Although both PSSK and TCBK demonstrate strong performance under controlled conditions, several limitations must be acknowledged. First, the relationship between HR and CBT is highly state-dependent. Even though HR is a useful proxy signal, its correlation with CT varies significantly across activity states, and any method that assumes a static HR–CBT relationship may suffer from systematic estimation bias. This issue is particularly important for HR-only models and necessitates careful modeling of context-specific dynamics, as addressed by PSSK.

Second, while PSSK effectively models such dynamics, it relies on accurate physiological state labels for training and deployment. In real-world applications, these labels are often unavailable or unreliable. Although the TCBK method avoids this requirement by clustering entire trials, it is more susceptible to overfitting to dataset-specific trends and shows weaker generalizability when applied to unseen datasets.

Third, both datasets used in this study feature participants with relatively homogeneous age distributions and were collected under limited environmental variation. This may partially explain the high within-dataset performance of both models and raises concerns about their applicability to broader and more diverse populations.

Another practical limitation arises from the initialization requirement of KF-based methods. Accurate CBT estimation depends on the initial CT value at $t = 0$. However, in realistic deployment scenarios, such accurate initialization is often unavailable. In this study, we set the initial CT to 37.0°C with a initial variance of 0.01. Although this approximation is sufficient for experimental consistency, providing a more accurate initial CBT–e.g. via short-term calibration or one-time measurement–could further improve estimation performance in practical applications.

In addition, following the SAGER guidelines, we examined whether model performance differed by sex. Using Dataset 1, which includes male and female participants, we compared the per-trial RMSEs for both models. For PSSK, the aggregated RMSEs were 0.40 for males and 0.43 for females ($p = 0.71$, Mann–Whitney U test). For TCBK, the respective RMSEs were 0.32 and 0.43 ($p = 0.21$). In both cases, $p > 0.05$, indicating no statistically significant performance differences between sexes. These findings suggest that the proposed models generalize well across male and female participants in this dataset.

In summary, although multimodal approaches incorporating additional sensors may offer potential accuracy gains, our findings suggest that HR-only models strike a favorable balance between simplicity, performance, and deployability. Future work should explore adaptive fusion strategies that combine HR with other physiological signals in a context-aware manner, while also validating the proposed models across broader demographic and environmental settings.

5. Conclusion

We proposed two KF-based models for CBT estimation using only HR: the PSSK and the TCBK filter. Both methods are lightweight and suitable for deployment on wearable devices.

PSSK incorporates physiological state labels to model state-dependent dynamics, while TCBK uses trial-level clustering based on HR and CBT statistics. Experiments on two datasets show that PSSK achieves robust performance across datasets, while TCBK performs best within datasets by capturing participant- and context-specific variability.

These results highlight the potential of HR-only models for accurate, efficient, and interpretable CBT estimation. Future work will explore adaptive state classification, multimodal integration, and large-scale validation in diverse populations and real-world environments.

Data availability statement

The implementation code for the proposed methods is available at: https://github.com/Oxford-NIL/PSSK-TCBK. Dataset 1 can be accessed from the Loughborough University Research Repository via: https://doi.org/10.17028/rd.lboro.26076577.v1. Dataset 2 is not publicly available due to usage restrictions but can be requested from the corresponding author of the original publication. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC), grant number UKRI012, and by the National Institute for Health and care Research (NIHR) HealthTech Research Centre for Community Healthcare at Oxford Health NHS Foundation Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Conflict of interest

The authors declare no conflict of interest.

ORCID iDs

Yuanzhe Zhao 0009-0008-5159-0940
Jeroen HM Bergmann 0000-0001-7306-2630

References

Arbury S, Jacklitsch B, Farquah O, Hodgson M, Lamson G, Martin H and Profitt A 2014 Heat illness and death among workers-United States, 2012–2013 *Morb. Mortal. Wkly. Rep.* **63** 661–5

Buller M J, Delves S K, Fogarty A L and Veenstra B J 2021 On the real-time prevention and monitoring of exertional heat illness in military personnel *J. Sci. Med. Sport* **24** 975–81

Buller M J, Tharion W J, Cheuvront S N, Montain S J, Kenefick R W, Castellani J and Hoyt R W 2013 Estimation of human core temperature from sequential heart rate observations *Physiol. Meas.* **34** 781

Buller M J, Tharion W J, Duhamel C M and Yokota M 2015 Real-time core body temperature estimation from heart rate for first responders wearing different levels of personal protective equipment *Ergonomics* **58** 1830–41

Casa D J, Becker S M, Ganio M S, Brown C M, Yeargin S W, Roti M W and Maresh C M 2007 Validity of devices that assess body temperature during outdoor exercise in the heat *J. Athl. Train.* **42** 333

Dolson C M, Harlow E R, Phelan D M, Gabbett T J, Gaal B, McMellen C and Seshadri D R 2022 Wearable sensor technology to predict core body temperature: a systematic review *Sensors* **22** 7639

Eggenberger P, MacRae B A, Kemp S, Bürgisser M, Rossi R M and Annaheim S 2018 Prediction of core body temperature based on skin temperature, heat flux and heart rate under different exercise and clothing conditions in the heat in young adult males *Front. Physiol.* **9** 1780

Falcone T, Del Ferraro S, Molinaro V, Zollo L and Lenzuni P 2024 A real-time biphasic Kalman filter-based model for estimating human core temperature from heart rate measurements for application in the occupational field *Front. Public Health* **12** 1219595

Faurie C, Varghese B M, Liu J and Bi P 2022 Association between high temperature and heatwaves with heat-related illnesses: a systematic review and meta-analysis *Sci. Total Environ.* **852** 158332

Fiala D, Lomas K J and Stohrer M 1999 A computer model of human thermoregulation for a wide range of environmental conditions: the passive system *J. Appl. Physiol.* **87** 1957–72

Fu M, Weng W, Chen W and Luo N 2016 Review on modeling heat transfer and thermoregulatory responses in human body *J. Therm. Biol.* **62** 189–200

Fuller D, Colwell E, Low J, Orychock K, Tobin M A, Simango B and Taylor N G 2020 Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure and heart rate: systematic review *JMIR mHealth uHealth* **8** e18694

Ganio M S, Brown C M, Casa D J, Becker S M, Yeargin S W, McDermott B P, Boots L M, Boyd P W, Armstrong L E and Maresh C M 2009 Validity and reliability of devices that assess body temperature during indoor exercise in the heat *J. Athl. Train.* **44** 124–35

Garzón-Villalba X P, Wu Y, Ashley C D and Bernard T E 2017 Ability to discriminate between sustainable and unsustainable heat stress exposures-Part 2: physiological indicators *Ann. Work Expo. Health* **61** 621–32

Gauer R and Meyers B K 2019 Heat-related illnesses *Am. Fam. Phys.* **99** 482–9

Giavarina D 2015 Understanding bland altman analysis *Biochem. Med.* **25** 141–51

Han X, Wu J, Hu Z, Li C and Hu X 2025 A practical deep learning model for core temperature prediction of specialized workers in high-temperature environments *J. Therm. Biol.* **128** 104079

Havenith G, Davey S, Downie V, Griggs K and Richmond V 2024 Dataset for 'Prediction of Core Body Temperature from Multiple Variables' and 'The physiological strain index does not reliably identify individuals at risk of reaching a thermal tolerance limit' PROSPIE project *Loughborough University* (https://doi.org/10.17028/rd.lboro.26076577.v1)

Hess J J, Saha S and Luber G 2014 Summertime acute heat illness in US emergency departments from 2006 through 2010: analysis of a nationally representative sample *Environ. Health Perspect.* **122** 1209–15

Hooper V D and Andrews J O 2006 Accuracy of noninvasive core temperature measurement in acutely ill adults: the state of the science *Biol. Res. Nurs.* **8** 24–34

Horn G P, Blevins S, Fernhall B and Smith D L 2013 Core temperature and heart rate response to repeated bouts of firefighting activities *Ergonomics* **56** 1465–73

Jardine D S 2007 Heat illness and heat stroke *Pediatr. Rev.* **28** 249–58

Karvonen J and Vuorimaa T 1988 Heart rate and exercise intensity during sports activities *Sports Med.* **5** 303–11

Khodarahmi M and Maihami V 2023 A review on Kalman filter models *Arch. Comput. Methods Eng.* **30** 727–47

Kubota N, Okada K and Yamanaka Y 2025 Circadian phase assessment of core body temperature using a wearable temperature sensor under the real world *Sleep Sci.* **18** 246–52

Laxminarayan S, Buller M J, Tharion W J and Reifman J 2014 Human core temperature prediction for heat-injury prevention *IEEE J. Biomed. Health Inform.* **19** 883–91

Moran D S and Mendal L 2002 Core temperature measurement: methods and current insights *Sports Med.* **32** 879–85

Moyen N E, Bapat R C, Tan B, Hunt L A, Jay O and Mündel T 2021 Accuracy of algorithm to non-invasively predict core body temperature using the Kenzen wearable device *Int. J. Environ. Res. Public Health* **18** 13126

Niedermann R *et al* 2014 Prediction of human core body temperature using non-invasive measurement methods *Int. J. Biometeorol.* **58** 7–15

Richmond V L, Davey S, Griggs K and Havenith G 2015 Prediction of core body temperature from multiple variables *Ann. Occup. Hyg.* **59** 1168–78

Rizvi I H 2022 A modified Kalman filter-based model for core temperature estimation during exercise and recovery with/without personal cooling interventions *J. Therm. Biol.* **109** 103307

Rowell L B 1974 Human cardiovascular adjustments to exercise and thermal stress *Physiol. Rev.* **54** 75–159

Rubin S A 1987 Core temperature regulation of heart rate during exercise in humans *J. Appl. Physiol.* **62** 1997–2002

Soehle M, Dehne H, Hoeft A and Zenker S 2020 Accuracy of the non-invasive Tcorem temperature monitoring system to measure body core temperature in abdominal surgery *J. Clin. Monit. Comput.* **34** 1361–7

Stolwijk J A 1971 A mathematical model of physiological temperature regulation in man *NASA Contractor Rep.* CR-1855

Tokizawa K, Shimuta T and Tsuchimoto H 2022 Validity of a wearable core temperature estimation system in heat using patch-type sensors on the chest *J. Therm. Biol.* **108** 103294

Venugopal V, Rekha S, Manikandan K, Latha P K, Vennila V, Ganesan N and Chinnadurai S J 2016 Heat stress and inadequate sanitary facilities at workplaces-an occupational health concern for women? *Glob. Health Action* **9** 31945

Verdel N, Podlogar T, Ciuha U, Holmberg H-C, Debevec T and Supej M 2021 Reliability and validity of the CORE sensor to assess core body temperature during cycling exercise *Sensors* **21** 5932

Welles A P, Buller M J, Looney D P, Rumpler W V, Gribok A V and Hoyt R W 2018 Estimation of metabolic energy expenditure from core temperature using a human thermoregulatory model *J. Therm. Biol.* **72** 44–52

Welles A P, Xu X, Santee W R, Looney D P, Buller M J, Potter A W and Hoyt R W 2018 Estimation of core body temperature from skin temperature, heat flux and heart rate using a Kalman filter *Comput. Biol. Med.* **99** 1–6

Wyss C R, Brengelmann G L, Johnson J M, Rowell L B and Niederberger M 1974 Control of skin blood flow, sweating and heart rate: role of skin vs. core temperature *J. Appl. Physiol.* **36** 726–33

Xu X and Werner J 1997 A dynamic model of the human/clothing/environment-system *Appl. Hum. Sci.* **16** 61–75

Yang J, Weng W, Wang F and Song G 2017 Integrating a human thermoregulatory model with a clothing model to predict core and skin temperatures *Appl. Ergon.* **61** 168–77

Yokota M, Berglund L G, Santee W R, Buller M J, Karis A J, Roberts W S, Cuddy J S, Ruby B C and Hoyt R W 2012 Applications of real-time thermoregulatory models to occupational heat stress: validation with military and civilian field studies *J. Strength Cond. Res.* **26** S37–44

Zhang W, Li L, Wang Y, Dong X, Liu C, Sun L and Xu S 2025 Continuous core body temperature monitoring for heatstroke alert via a wearable in-ear thermometer *ACS Sens.* **10** 1440–9

Zhao Y and Bergmann J H M 2023 Non-contact infrared thermometers and thermal scanners for human body temperature monitoring: a systematic review *Sensors* **23** 7439

Zhao Y and Bergmann J 2025 A hybrid core temperature estimation method integrating physiological and environmental factors *IEEE Sens. Lett.* **9** 1–4

Zhao Y, de Almeida e Bueno L, Holdsworth D A and Bergmann J H M 2024 Evaluating the agreement between oral, armpit and ear temperature readings during physical activities in an outdoor setting *Int. J. Environ. Res. Public Health* **21** 595