A method of evaluating the accuracy of human body thermoregulation models

Article

Accepted Version

Building and Environment 87 (2015) 1–9

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv

A method of evaluating the accuracy of human body thermoregulation models

Yu Yang [a, b, c], Runming Yao [c, b], Baizhan Li [a, b, *], Hong Liu [a, b], Lai Jiang [c]

[a] Key Laboratory of the Three Gorges Reservoir Region's Eco-Environment, Ministry of Education, Chongqing University, Chongqing 400045, China
[b] National Centre for International Research of Low-carbon and Green Buildings, Chongqing University, Chongqing 400045, China
[c] School of Construction Management and Engineering, University of Reading, UK

A Method of Evaluating the Accuracy of Human Body Thermoregulation Models

Yu Yang [a,b,c]; Runming Yao [c,b]; Baizhan Li [a,b*]; Hong Liu [a,b]; Lai Jiang [c]

[a] Key Laboratory of the Three Gorges Reservoir Region's Eco-Environment, Ministry of Education, Chongqing University, Chongqing 400045, China;

[b] National Centre for International Research of Low-carbon and Green Buildings, Chongqing University, Chongqing 400045, China;

[c] School of Construction Management and Engineering, University of Reading

Highlights:

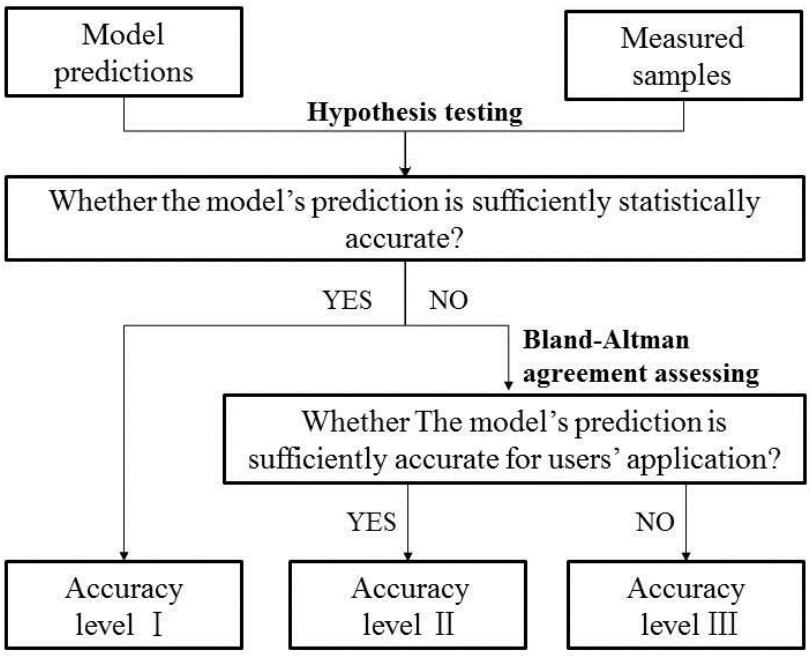
- Lack of studies into evaluating population-based Human Body Thermoregulation models;
- A new evaluation method of combined both statistical and empirical methods;
- Evaluation of the population-based HBT models' accuracy;
- A framework for the validation process of HBT models.

Abstract

Human Body Thermoregulation Models have been widely used in the field of human physiology or thermal comfort studies. However there are few studies on the evaluation

method for these models. This paper summarises the existing evaluation methods and critically analyses the flaws. Based on that, a method for the evaluating the accuracy of the Human Body Thermoregulation models is proposed. The new evaluation method contributes to the development of Human Body Thermoregulation models and validates their accuracy both statistically and empirically. The accuracy of different models can be compared by the new method. Furthermore, the new method is not only suitable for the evaluation of Human Body Thermoregulation Models, but also can be theoretically applied to the evaluation of the accuracy of the population-based models in other research fields.

Graphic Abstract



Keywords

Evaluation Method; Model Evaluation; Prediction Accuracy; Indoor environment; Thermal Comfort.

Nomenclature

ave_i	mean value of $\mu_{m,i}$ and x_i	s_i^2	sample variances from population 'i'
CI	confidence interval for population mean	$\bar{T}$	sample mean of skin temperature
$\bar{d}$	mean difference between sample means and model predictions	T	skin temperature
d_i	difference between sample mean and model prediction of population 'i'	X_i	population 'i'
H0	null hypothesis	$\bar{x_i}$	sample mean from population 'i'
HBT	human body thermoregulation	$x_{i,j}$	sample j from population 'i'
k	number of the populations	α	significance level
LOA	limit of agreement	μ_a	prediction from Model A
n_i	number of the samples from population 'i'	μ_b	prediction from Model B
N_{ob}	number of the observations	μ_m	model prediction
N_{sub}	number of the subjects	μ_i	population mean
RMSE	root mean square error		
s_d	standard deviation of the differences between sample means and model predictions	Subscript	
se_i	standard error of sample mean from population 'i'	a	denotes Model A
		b	denotes Model B
		i	denotes population number
		j	denotes sample number
		m	denotes model
		t	denotes time

1. Introduction

The thermal interaction of the human body with the environment involves two processes: i) the heat transfer between the human body and the thermal environment, simultaneously including radiation, convection, conduction, evaporation and respiration; and ii) the self-regulation function of the human body which responds to varied thermal environments, such as vasoconstriction, vasodilation, shivering and sweating [1]. Human Body Thermoregulation Models (HBT models) are developed to simulate these two aspects of interaction and then predict the human thermal physiological responses (e.g. skin temperature, core temperature) under thermal conditions usually with the input parameters of air temperature, radiation temperature, air velocity, relative humidity, clothing insulation, metabolic rate and their variations with exposure time. These models have been widely used in the field of human physiology or thermal comfort studies.

The existing research in this field mainly focuses on developing HBT models using different modelling methods for body segmentation [2-6], thermoregulatory systems [2, 7, 8], heat transfer [3, 5] and numerical solutions [3, 9]. It is very important to evaluate the accuracy of the models. However, very little effort has been made to study the methods for evaluating the prediction accuracy of the HBT models. It is still a question under discussion whether the existing model-evaluation methods are generally applicable.

Models predicting the average thermal responses of a group of human bodies are defined as ‘population based’ model, and this average response is recognized as the ‘population mean’ in statistics. The existing HBT models are mostly population-based because individual thermal responses vary from one person to another. Two questions in evaluating the prediction accuracy of HBT models are still open: i) How to validate the prediction accuracy of the models in given thermal processes. This is because the users need to have confidence in applying the models in practice. And ii) How to compare the prediction accuracy of models for the same thermal processes. This is to provide guidance for the selection of the most accurate one among different models.

In this paper, the existing evaluation methods for HBT models are summarized and the strengths/weaknesses of these methods are analysed. Thereafter, a new evaluation method for HBT models has been developed.

2. Existing methods for evaluating the accuracy of HBT models

2.1 Brief literature review

This study has reviewed the accessible research papers published over the last fifty years in regard to the development or improvement of HBT models. In total, twenty-two related papers were selected for the discussion in this paper. The detailed information of model evaluation and evaluation methods in these studies is listed in Table 1. From the table, we can see that among these studies on the HBT models development, four papers proposed models without any evaluation; eighteen papers validated the prediction accuracy of the models and eight papers compared the prediction accuracy of different models.

The methods for evaluating models' accuracy in these papers can be summarized into two categories: i) directly observing the figures by comparing the predicted values from the models with the raw data or descriptive statistics of samples from experiments; which can be termed an 'Observation Method'; and ii) calculating the root mean square error (RMSE) between the model predictions and sample means; hereafter known as the 'RMSE Method'. From Table 1 we can see that fifteen papers utilised the 'observation' method and three papers applied the 'RMSE' method. Among the eight papers that compared the accuracy of different models; six used the 'observation' method and two used the 'RMSE' method.

Table 1. The methods used to evaluate from existing HBT models papers

No.	Model Reference	Whether the study validated the models' prediction accuracy		Whether the study compared the prediction accuracy of different models	
		Yes (Y) or No (N)	Method used	Yes (Y) or No (N)	Method used

1	[2]	N	N/A	N	N/A
2	[3]	Y	Observation	N	N/A
3	[10]	Y	Observation	N	N/A
4	[11]	Y	Observation	N	N/A
5	[12]	N	N/A	N	N/A
6	[13]	Y	RMSE	N	N/A
7	[14]	N	N/A	N	N/A
8	[15]	Y	Observation	N	N/A
9	[9]	Y	Observation	N	N/A
10	[8]	Y	RMSE	Y	Observation
11	[5]	Y	Observation	N	N/A
12	[4]	N	N/A	N	N/A
13	[16]	Y	Observation	N	N/A
14	[17]	Y	Observation	Y	Observation
15	[18]	Y	Observation	Y	Observation
16	[19]	Y	Observation	Y	Observation
17	[20]	Y	Observation	Y	Observation
18	[21]	Y	Observation	N	N/A
19	[7]	Y	Observation	Y	RMSE
20	[22]	Y	Observation	N	N/A
21	[6]	Y	Observation	Y	Observation
22	[23]	Y	RMSE	Y	RMSE

2.2 Analysis of the existing methods

The ‘Observation’ and ‘RMSE’ methods, to some extent, are insufficient to evaluate the prediction accuracy of the HBT models. We use a practical example of real data from our experimental studies for a further explanation (see in Figure 1).

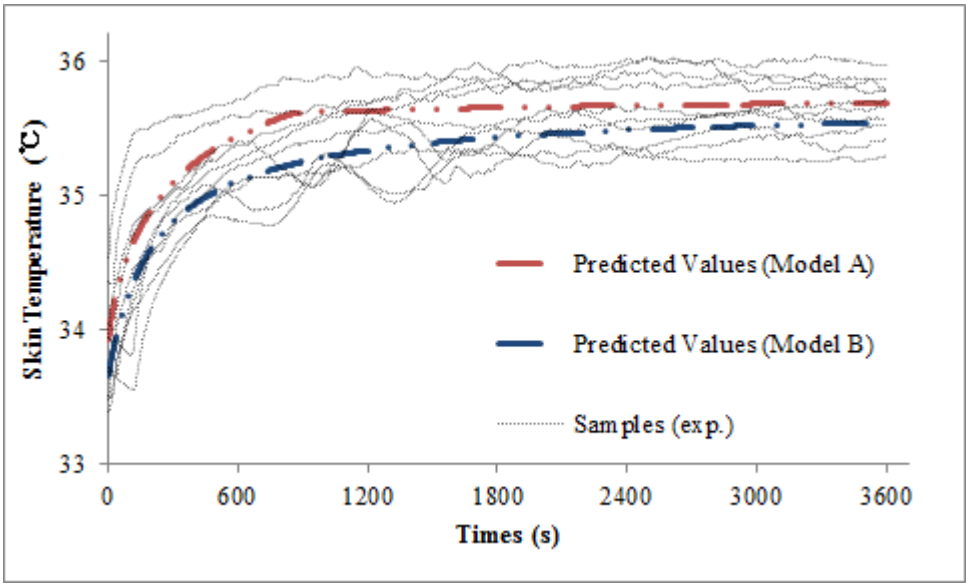


Figure 1. The raw dataset of predictions and samples

These black thin lines in Figure 1 show the raw data of measurements for skin temperatures in a human exposure experiment in which ten half-naked health male subjects experienced a temperature step-change process from The Environment Ⅰ to the Environment Ⅱ and then stayed in the Environment Ⅱ for a period of 3600 seconds. The information of the subjects and the thermal conditions of the two environments are listed in Table 2 and 3 respectively. Subjects' skin temperatures were measured each second. The measured skin temperature of subject '*j*' at time '*t*' is expressed as $T_{t,j}$ (t=1 ••• 3600; j=1 ••• 10).

Table 2. Subjects' information in the case study (mean ± standard deviation)

Age	24±1
Height (m)	174±6
Weight (kg)	60±7
Clothing insulation (clo)	0.03±0
Activity level (met)	1.0±0

Table 3. Thermal conditions of the experiment in case study (mean ± standard deviation)

	Environment Ⅰ	Environment Ⅱ
Air Temperature (℃)	28.2±0.1	34.7±0.1
Relative Humidity (%)	60.4±2.8	55.8±0.3
Air Velocity (m/s)	0.06±0.01	0.18±0.04
Globe Temperature (℃)	28.3±0.25	34.7±0.1

Models A and B are two modified HBT models based on the classical two-node model of thermoregulation [2]. The two models were developed by optimizing the modelling of the body, the regulatory system and the numerical solution method. The main difference between the two models is the empirical parameters of the regulatory system (i.e. function of regulatory sweating rate and blood flow rate of skin), which are achieved by training different sets of data, respectively. Both of the models are applied to simulate the skin temperature for the above thermal process. The predicted data from Model A and Model B for each second are denoted as $\mu_{a,t}$, $\mu_{b,t}$ ($t=1,2,3 \cdots 3600$), which are represented by red and blue lines in Figure 1.

Here we attempt to use these available data ($T_{t,j}$, $\mu_{a,t}$, $\mu_{b,t}$) to validate and compare the prediction accuracy of these two models for this specific thermal process. The existing methods for evaluating models' prediction accuracy are analysed using this example.

When applying the 'Observation Method' or 'RMSE Method', usually the first step is to calculate the average skin temperature of the 10 subjects at each moment t (t=1 $\cdots$ 3600) by $T_{t,j}$. This is generally known as the sample mean of skin temperature which is expressed as $\overline{T}_t$ in equation 1.

$$\overline{T}_t = \frac{1}{N_{sub}} \sum_{j=1}^{N_{sub}} T_{t,j} \qquad \textit{Equation 1}$$

In the 'Observation Method', the most common way is to depict the sample means and model predictions in a figure and draw conclusions concerning the prediction accuracy of models through observing the relationships between the data in the figure. In this

example, every second the measured sample mean and the predictions from Model A and Model B are plotted in Figure 2, and it is the most typical figure that appears in papers using the 'Observation Method' (e.g. reference [6]). It can be found that predictions of Model A in the first 1800 seconds are about 0.05-0.3℃ higher than the sample means; the differences between the predictions and sample means in the final 1800s are less than 0.05℃; the maximum deviation of these two sets of data is about 0.3℃. By contrast, predictions of Model B are always about 0.1-0.2℃ less than the sample means. By investigating the information obtained from Figure 2, we can neither draw conclusion on whether the accuracy of Model A or Model B is sufficiently accurate in predicting or on which model predicts more accurately than the other. Therefore, in this example, the 'Observation Methods' are not applicable for the evaluation of HBT models.

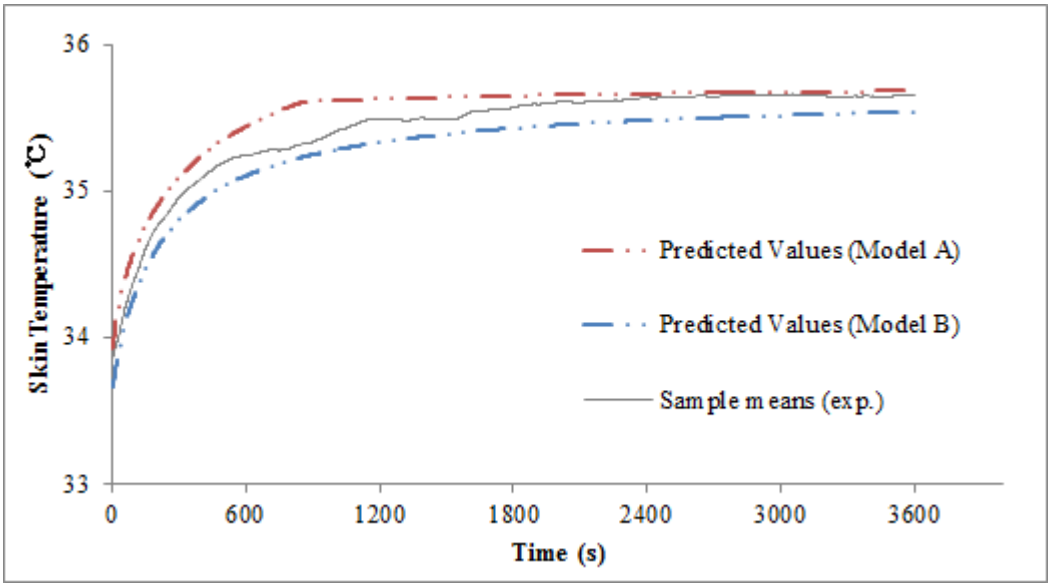


Figure 2. Model predictions and measured sample means

The 'RMSE Method' is applied based on the RMSE values between model predictions and sample means, which can be generally expressed as Equation 2 . In this example, RMSE values ($RMSE_a$, $RMSE_b$) of Model A and Model B has been calculated as 0.13 and 0.136 respectively for evaluating the models' accuracy. As the 'RMSE Method' is applied in reference [23], the conclusion about whether the prediction of Model A or

Model B is sufficiently accurate or not should directly depend on the RMSE value of 0.13 or 0.136. In addition, a conclusion that Model A is better than Model B in prediction accuracy will be drawn according to the fact that $RMSE_a < RMSE_b$ as described in reference [7].

$$RMSE_m = \sqrt{\frac{1}{N_{ob}} \sum_{t=1}^{N_{ob}} \left(\mu_{m,t} - \overline{T}_t \right)^2} \qquad \textit{Equation 2}$$

However, the existing references for HBT models did not provide any theoretical basis or statistical reference for the 'RMSE Method' to support the conclusions. Actually, RMSE is one of the most popular error measures of prediction accuracy [24], but it is commonly used for comparing the accuracy of models [24] rather than validating the accuracy of a model, because i) RMSE index itself cannot be used for statistical inference on validation; and ii) itself lacks of specific meaning for common user to understand the accuracy of the model. Besides, the error measure of RMSE is not appropriate for all the accuracy comparison studies as it also has its constraints [24]. A recognized constraint of RMSE when applied to HBT models is the ignorance of the difference between the populations and samples, which will be discussed in Section 3.1.1.

By analysing the existing methods for evaluating the prediction accuracy of HBT models, we can summarize that: i) the 'Observation Method' is a simple and straightforward decision-making method but in many cases it cannot provide a convincing evaluation of models because it lacks a theoretical basis; ii) the RMSE is a useful measure of accuracy but can only reasonably be applied in comparing the prediction accuracy of models with some constraints specified for population-based HBT models. Considering the inadequacy of existing methods, a general method for evaluating the accuracy of HBT models is necessary.

3. The evaluation method

3.1 Principles

As stated above, to evaluate the accuracy of HBT models is to i) validate and ii) compare their prediction accuracy. The proposed evaluation method attempts to solve these two problems separately.

3.1.1 Validate the prediction accuracy of models

A requirement of the developed models is that they can accurately predict the real situation. But how can we judge the accuracy of the model? The proposed method considers this question from the following two aspects:

1) Statistical Analysis

As the HBT model is a population-based model of predicting population means, therefore, theoretically, the accuracy evaluation should be based on the measurements of the differences between the predictions from the model and the population means. However, population means are usually unknown and often unavailable. Most existing studies use sample means instead of population means to calculate the discrepancies because the sample mean is the unbiased estimator of the population mean. When the sample size is small or the sample variation is large, both of which are very common in the existing human thermal response studies, the variance of the unbiased estimator will be large and the sample mean might often be far from the true mean. If the sample mean is used instead of the population mean, the true value might be distorted due to the lack of statistical information such as sample size and sample standard deviation. In fact, when statistically estimating the population mean, interval estimation which describes the population mean using confidence interval consisting of the sample mean and standard deviation, is a more scientific approach than point estimation which characterizes the population mean via the sample mean. The confidence interval provides a range that is highly likely (often 95% or 99%) to contain the true population quantity that is being estimated, and through which the researchers can analyse the

difference between the prediction and population mean by statistical inference. In this way, the accuracy of the model can be validated statistically.

2) Empirical Analysis

Apart from analysis of inferential statistics, the degree of agreement between model predictions and sample means can equally reveal the degree of accuracy of the model. Appropriate methods for assessing the agreement between model predictions and sample means are needed in a model evaluation process.

Bland and Altman proposed a method to assess agreement between two measurement methods in clinical research. They criticized the commonly-used approaches including 'Comparison of means', 'Correlation coefficient', and 'Regression' as inappropriate ways for the agreement assessment of two different measures [25] and proposed a new approach which was named *the Bland-Altman analysis* [26, 27]. Zaki [28] endorsed that in medical research, the Bland-Altman method is the most popular method for agreement studies; 85% studies having used this method during 2007 to 2009. In this paper, in order to assess the agreement between the sample means and predictions of the HBT model, we introduce the Bland-Altman method. The sample means and the predictions from the HBT model can be regarded as two methods for measuring the population means. To apply the Bland-Altman method, we calculate the mean difference $(\bar{d})$ of the level of population means obtained by sample means and model predictions, and also calculate the standard deviation of the differences (s_d). Consequently, the index 'limits of agreement' ($\bar{d} \pm 1.96 s_d$), which represents the range in which 95% differences between the predictions and the sample means will lie, is obtained. Consequently, the degree of agreement between the sample means and the model predictions is dependent on 'whether the differences provided by the 'limits of agreement' are acceptable by the users in application'. By agreement analysis, the accuracy of the model can be validated empirically.

3.1.2 Comparisons of the prediction accuracy of models

In the study of developing HBT models, it is common to compare the accuracy of different models using the samples from the same dataset to select the model with the better/best accuracy. RMSE is a commonly-used error analysis measure for comparing the prediction accuracy of models, but it has some constraints when applied to certain models. RMSE represents the average closeness of the predicted data to the 'sample means' but not to the 'population means'. According to the aforementioned analysis of the difference between the population means and the sample means, a remarkable inadequacy of the traditional 'RMSE Method' is that it ignores the analysis on populations when comparing models. As the relationship between the predictions and populations has already been analyzed in the statistical validation process of section 3.1.1, applying index RMSE based on the results of models' validation will be an improvement over the traditional RMSE method by taking the factor of population into consideration when comparing the prediction accuracy of models.

Models whose accuracy is validated statistically are more acceptably accurate than models whose accuracy is validated empirically. If the models are validated as the same accuracy level (see in section 3.2), the RMSE values calculated from the models' predictions and sample means are applied for the further comparison, considering that the statistical validation process is completely objective while the empirical validations are subjective .

3.2 The process of evaluation

Based on the above discussion, a new method for the accuracy evaluation of HBT models is proposed here:

Set of data:

The population *"i"* is denoted by X_i which represents a physiological index set (such as skin temperature, core temperature, etc.) under certain conditions. Its mean is denoted by μ_i. In the conditions for evaluation, the HBT model *m* is used to predict the population means $\mu_i (i=1 \bullet\bullet\bullet k)$ from *k* populations $X_i (i=1 \bullet\bullet\bullet k)$ and the prediction for each population is described as $\mu_{m,i}$ $(i=1 \bullet\bullet\bullet k)$; $x_{i,j}$ $(j=1 \bullet\bullet\bullet n_i)$ are n_i samples

from X_i. The sample mean, sample variance and standard error of sample mean from X_i are denoted as $\bar{x}_i$, s_i^2 and se_i respectively. In the statistical analysis, the significance level is denoted as α and the value chosen in this study is α=0.05.

Evaluation steps:

1) Define the null hypothesis and alternative hypothesis.

The null hypothesis is "For any populations provided, the model can accurately predict the population mean, that is $\mu_{m,i} = \mu_i$ (i=1 ••• k)'. The alternative hypothesis is "the model cannot accurately predict all the population means, that is at least for one value of i, $\mu_{m,i} \neq \mu_i$".

2) Define the confidence intervals for population means.

By calculating $\bar{x}_i$ (Equation 3), s_i (Equation 4), se_i (Equation 5) of each population, the $100(1-\alpha)$% confidence interval CI_i for each population mean can be derived (Equation 6) [29]. The probability that CI_i contains μ_i is $100(1-\alpha)\%$.

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} \qquad \textit{Equation 3}$$

$$s_i = \sqrt{s_i^2} = \sqrt{\frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}_i\right)^2} \qquad \textit{Equation 4}$$

$$se_i = \frac{s_i}{\sqrt{n_i}} \qquad \textit{Equation 5}$$

$$CI_i = \left[\bar{x}_i - t_{\alpha/2}(n_i - 1)\,se_i, \bar{x}_i + t_{\alpha/2}(n_i - 1)\,se_i\right] \qquad \textit{Equation 6}$$

In Equation 6, $t_{\alpha/2}(n_i - 1)$ is the $\alpha/2$ percentage point of the t-distribution with $(n_i - 1)$ degrees of freedom, which is determined by the value of α and ($n_i - 1$). For the $100(1-\alpha)\%$=95% confidence interval, $t_{\alpha/2}(n_i - 1)$ values are 2.2, 2.1 and 2.0 for 10, 20 and 30 degrees of freedom respectively.

3) Validate the model's accuracy by comparing model predictions with the confidence intervals of the population means.

Compare all the predictions $\mu_{m,i}$ with the corresponding confidence interval CI_i (this can be conveniently judged by the graphical method): according to statistical inference, if all the model predictions are within the corresponding confidence interval, it indicates that the difference between predictions and population means are statistically insignificant (α =0.05). Therefore, the hypothesis $H0$ cannot be rejected which suggests the model's prediction is statistically accurate. The model in this case will be classified into accuracy level Ⅰ. On the other hand, if one or more of the predictions fall outside the corresponding confidence interval, it indicates that at least one of the model predictions is significantly statistically different from the population mean (α =0.05). Therefore, the hypothesis $H0$ should be rejected, and the model's accuracy cannot be validated by statistical inference.

A model of accuracy level I means: through statistical inference, the model is validated to be sufficiently accurate for predicting, i.e. "the model's prediction is statistically accurate". For models which are not classified into level Ⅰ, the accuracy fails to satisfy the statistical requirements. Hence, further empirical validation is needed.

4) Validate the model's accuracy by analysing the agreement between the model prediction and the sample mean.

For models that cannot be validated statistically, the Bland-Altman method [30] of agreement study is introduced to analyse the agreement between the model prediction and sample mean, based on which, the empirical validation can be made from the requirements of the model application. The specific application of the Bland-Altman method in a model evaluation study is as follows: for the population i, d_i is the differences between $\mu_{m,i}$ and $\bar{x}_i$ (Equation 7) and ave_i is the mean value of $\mu_{m,i}$ and $\bar{x}_i$ (Equation 8); for the differences from all the populations, the mean $\bar{d}$ and standard deviation s_d (Equation 10) of these differences can be calculated in Equation

9 and Equation 10, respectively; If the differences are normally distributed, we would expect 95% of them to lie within $\bar{d} \pm 1.96 s_d$, which we call 95% limits of agreement. For any α, the 'limits of agreement' (*LOA*) of 1-α can be calculated as shown in Equation (12). These values define the range within which most differences between the predictions and the sample means will lie. The decision on the accuracy of the model is made by the users. If the user considers that the difference provided by the 'limit of agreement' is acceptable when applying the model, it suggests that the predictions have good agreement with the sample means. In this case, we regard the model's accuracy as level Ⅱ, which means "the model's prediction is sufficiently accurate for users in application". Conversely, if the user judges the provided difference between the sample mean and the prediction as significant for the model application and cannot accept it, the model's accuracy will be classified into level Ⅲ, which implies "the model's prediction is not sufficiently accurate".

$$d_i = \mu_{m,i} - \bar{x}_i \qquad \textit{Equation 7}$$

$$ave_i = \frac{1}{2}\left(\mu_{m,i} + \bar{x}_i\right) \qquad \textit{Equation 8}$$

$$\bar{d} = \frac{1}{k}\sum_{i=1}^{k} d_i \qquad \textit{Equation 9}$$

$$s_d = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k} d_i} \qquad \textit{Equation 10}$$

$$LOA = \left[\bar{d} - z_{\alpha/2} s_d, \bar{d} + z_{\alpha/2} s_d\right] \qquad \textit{Equation 11}$$

In Equation 11, $z_{\alpha/2}$ is the $\alpha/2$-percentile of the standard normal distribution. When $\alpha = 0.05$, $z_{\alpha/2}$=1.96.

The stated calculation of the 'limits of agreement' is based on the assumption that differences are normally distributed. When differences do not follow normal distribution, the reference [26] indicated that 'a non-normal distribution of differences may not be as serious in Bland-Altman analysis as in other statistical contexts'. For example, for the 95% "limits of agreement", approximate analysis can still proceed as if the differences are normally distributed as long as 95% of the observed values of the difference lie within the intervals $\bar{d} \pm 1.96 s_d$. For the cases that are not in this scenario, reference [27] points out that 'this is perhaps most likely to happen when the difference and average value are related'. Considering that in this situation the calculation of the 'limits of agreement' will be complicated and this situation happens rarely, this part will not be elaborated in the present paper. Readers who are interested in this can directly refer to the paper [27].

It will be more convenient to use graphical techniques for the Bland-Altman analysis, which is described in the case study in section 4.

5) Compare the accuracy of models based on accuracy level and RMSE.

When different models applied to the same set of data, are compared the determination of accuracy should primarily depend on the models' accuracy level, and then be confirmed by comparing the RMSE (Equation 12) of the models. For these models, the accuracies of which are in different levels, a level I model is superior to the level Ⅱ model which is superior to the level Ⅲ model. When the models' accuracies are in the same level, the model with a smaller $RMSE$ is more accurate.

$$RMSE_m = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\mu_{m,i} - \bar{x}_i\right)^2} \qquad \textit{Equation 12}$$

In general, the first four steps show how to validate the prediction accuracy of the models and classify their accuracy level. The final step solves the problem of how to compare the prediction accuracy of the models, by which the more/most accurate model

can be selected. The models' validation and comparison process is summarized in Figure 3 and Table 4.

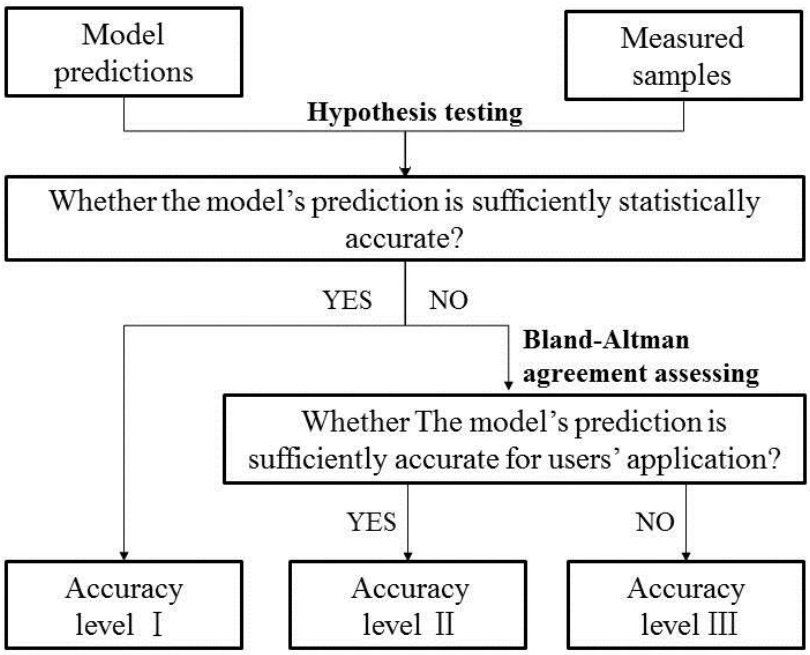


Figure 3. The framework for the validation process of HBT models

Table 4. Evaluation on the prediction accuracy of HBT models

Accuracy level	Term	Interpretation
Ⅰ	Statistically accurate	The model's prediction is sufficiently statistically accurate.
Ⅱ	Empirically accurate	The model's prediction is sufficiently accurate for users in application.
Ⅲ	Inaccurate	The model's prediction is not sufficiently accurate.

[a] For models in different accuracy levels, level Ⅰ models are more accurate than level Ⅱ models which are themselves more accurate than level Ⅲ models.

[b] For models in the same accuracy level, the smaller the model's ***RMSE*** is, the more accurate the model is.

4. Case study

In order to further explain this new evaluation method, a case study is illustrated here. The accuracy of Model A and Model B, which have been described in Section 2.2, will be evaluated by the new method.

First, the set of data in the case study needs to be linked to the corresponding inputs in the new method: these two HBT models are used to predict skin temperature per second in a given thermal process, the population *i* (X_i) is the set of skin temperature at the time *t* (that is *i* is equivalent to *t* in this case), thus the total numbers of the population are k=3600; the sample $x_{i,j}$ is the measured skin temperature $T_{t,j}$ and the sample size of each population is n_i=10; the predicted values of skin temperature per second from Model A or Model B are the predictions of population $\mu_{m,i}$ in the evaluation method, that is when evaluating Model A, $\mu_{m,i}$ is $\mu_{a,i}$ (i=1 ••• 3600), while when evaluating Model B, $\mu_{m,i}$ is $\mu_{b,i}$ (i=1 ••• 3600). In the case study, the significance level α is 0.05.

According to the new method, the evaluation process has five steps:

1) Define the null hypothesis $H0$ for Model A and Model B. The $H0$ for Model A (or Model B) is that 'the Model A (or Model B) can accurately predict population means. That is, for any i (i=1⋯3600), $\mu_{a,i} = \mu_i$ (or $\mu_{b,i} = \mu_i$)'.

2) Calculate the $\bar{x}_i$ (Equation 13), s_i (Equation 14), se_i (Equation 15) and build $100(1-\alpha)\% = 95\%$ CI_i (Equation 16) for each population. The calculated CI_i, $\bar{x}_i$ and $\mu_{a,i}$, $\mu_{b,i}$ (i=1⋯3600) are plotted in Figure 4.

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} = \frac{1}{10}\sum_{j=1}^{10} T_{i,\,j} \qquad \textit{Equation 13}$$

$$s_i = \sqrt{s_i^2} = \sqrt{\frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}_i\right)^2} = \sqrt{\frac{1}{9}\sum_{j=1}^{9}\left(T_{i,j} - \bar{x}_i\right)^2} \qquad \textit{Equation 14}$$

$$se_i = \frac{s_i}{\sqrt{n_i}} = \frac{s_i}{3} \qquad \textit{Equation 15}$$

$$CI_i = \left[\bar{x}_i - t_{\alpha/2}\left(n_i - 1\right)se_i, \bar{x}_i + t_{\alpha/2}\left(n_i - 1\right)se_i\right] = \left[\bar{x}_i - 2.2se_i, \bar{x}_i + 2.2se_i\right] \qquad \textit{Equation 16}$$

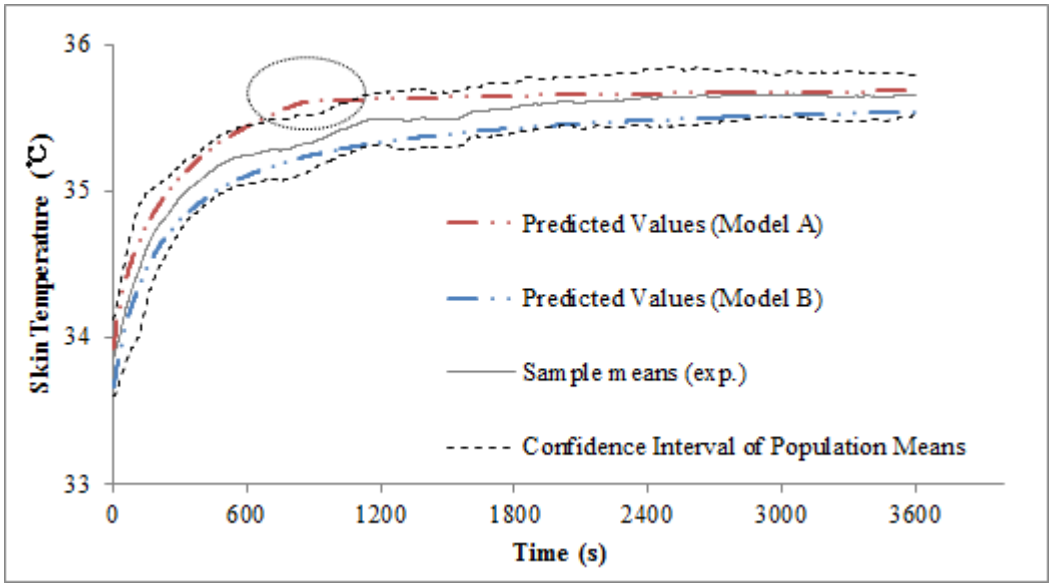


Figure 4. The predictions and statistical information from the samples

3) Compare the relationship between model predictions and the confidence intervals of the corresponding population. From Figure 4 it can be seen that for Model B, all predictions lie within the confidence intervals of the population means, indicating that there is no statistically significant difference between the population means and the predictions of Model B. Thus, the accuracy of Model B is evaluated as level Ⅰ. For Model A, the predictions for some populations are beyond their confidence intervals (marked in Figure 4), indicating that for these populations, the model's predictions have a statistically significant difference from the population means, so the model's accuracy

cannot be validated statistically. Further empirical validation for Model A should be carried out.

4) For Model A, the Bland-Altman method is applied to analyse the agreement of predictions and sample means. The statistics d_i, ave_i, $\bar{d}$, s_d are calculated from Equations 17 ~ 20. Dots (d_i, ave_i) are plotted in Figure 5, while the value of $\bar{d}$ and $\bar{d} \pm 1.96 s_d$ are marked in the figure. Figure 5 shows that there is no obvious correlation between the difference and average value, and 95% of the dots are located within the range $\bar{d} \pm 1.96 s_d$. Therefore, for Model B, the 'limit of agreement' of the model predictions and population means is -0.05 to 0.26 (Equation 21). Provided that the user regards the accuracy requirement for the skin temperature as ‘the difference between the model prediction and the sample mean in most cases must be less than 0.2℃ ’, due to the difference provided by the 'limit of agreement' having exceeded 0.2℃, the model’s accuracy will be evaluated as level Ⅲ.

$$d_i = \mu_{m,i} - \bar{x}_i = \mu_{b,i} - \bar{x}_i \qquad \textit{Equation 17}$$

$$ave_i = \frac{1}{2}\left(\mu_{m,i} + \bar{x}_i\right) = \frac{1}{2}\left(\mu_{b,i} + \bar{x}_i\right) \qquad \textit{Equation 18}$$

$$\bar{d} = \frac{1}{k}\sum_{i=1}^{k} d_i = \frac{1}{10}\sum_{i=1}^{10} d_i = 0.1 \qquad \textit{Equation 19}$$

$$s_d = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k} d_i} = \frac{1}{3}\sqrt{\sum_{i=1}^{10} d_i} = 0.08 \qquad \textit{Equation 20}$$

$$LOA = \left[\bar{d} - z_{\alpha/2} s_d, \bar{d} + z_{\alpha/2} s_d\right] = \left[\bar{d} - 1.96 s_d, \bar{d} + 1.96 s_d\right] = \left[-0.05, 0.26\right] \qquad \textit{Equation 21}$$

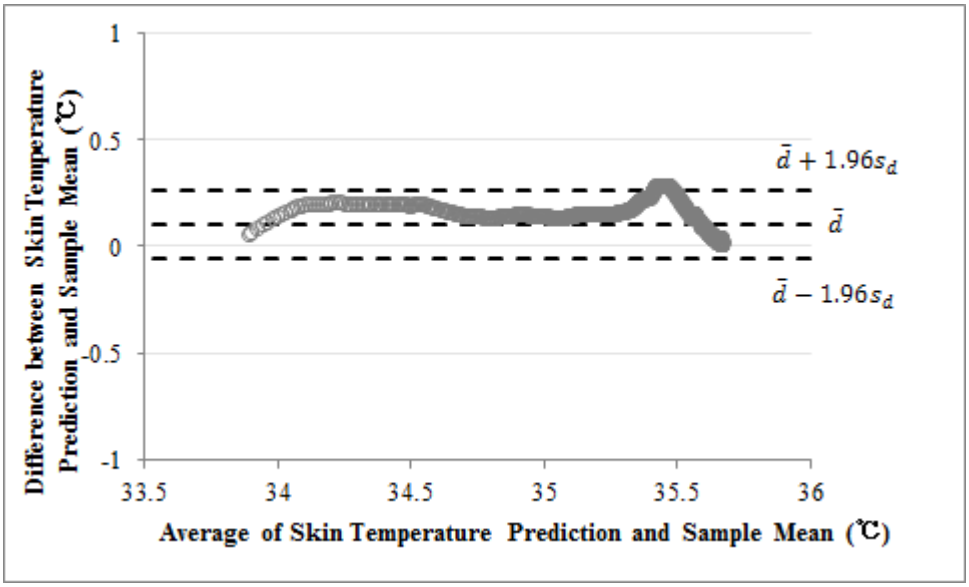


Figure 5. Bland- Altman analysis

5) Compare the accuracy between Model A and Model B. As the accuracy of Model A and Model B are evaluated as level Ⅲ and level Ⅰ respectively, the prediction accuracy of Model B is considered to be better than that of Model A.

The evaluation results of this case study can be concluded as follows: for the given thermal process, i) the accuracy of Model B is evaluated as level Ⅰ which means Model B's prediction is statistically accurate; ii) the accuracy of Model A is evaluated as level Ⅲ, thus, Model A is inaccurate in predicting the given process; iii) the prediction accuracy of Model B is better than that of Model A.

5. Discussion

5.1 The application of the new method

From the description of the new method and its application in the case study, some issues need to be pointed out when applying this new method:

1) The HBT models evaluated by this method should be population-based models, which are used to predict the average responses of populations. This method should not

be applied to any HBT model developed for individuals. Theoretically, the new evaluation method can be widely applied to any population-based model which includes, but is not limited to, the HBT models.

2) For every population predicted, a certain number of samples are required for a statistical validation process. According to statistical principles, there is no statistical approach that can give a 100% correct conclusion. When applying this method, the reliability of the conclusions increases with the sample size. Therefore, raw data with large sample size will be beneficial to the evaluation.

3) As the RMSE is the accuracy measure whose scale depends on the scale of the data [31], in the proposed evaluation method, the method of accuracy comparison of models is only applicable to situations in which models are applied to the same set of data. Models predicting different conditions are not comparable using this method.

4) In most of the HBT models studied, this method is used to evaluate models by examining the accuracy of the predictions. However, for some specialized models, the tendency of the predictions may be more important than the predicted values themselves. This method can be equally used for these models examining the accuracy of the changing rate of predictions - the principles are the same as when examining the accuracy of the predictions. Thus, when applying this evaluation method, the objects needing to be examined are dependent on the characteristics of the models.

In general, HBT models can easily satisfy these specificities mentioned above, which is the reason that this paper illustrates the new evaluation method through the example of HBT models. In theory, the application of this method can be extended to the evaluation of any models in other topics (such as the validation of thermal sensation models, as the questions be arisen in reference [32]).

5.2 The validation of models' prediction accuracy

From the case study, the prediction accuracy of models can be validated using both statistical and empirical analysis. The statistical validation only depends on the predictions from the model and the measured samples, thus the conclusion is objective.

The results of empirical validation are based on the set of data as well as the subjective judgments of the users. Thus, even provided with the same dataset, the conclusion may be different due to differences in users' requirements for accuracy. For example, in the case study, the accuracy of Model A cannot be statistically validated hence the empirical validation is used. The 'limit of agreement' is obtained as -0.05 to 0.26 through Bland-Altman analysis, but the user believes 'the bias between prediction and sample mean should not exceed 0.2℃', as the 'limit of agreement' is beyond this threshold of 0.2℃, the model is recognized as inaccurate for this thermal process. However, if for some reason, the user's requirement for accuracy becomes less rigorous and a bias which is less than 0.3℃ becomes acceptable, then Model A becomes sufficiently accurate for application by the users. Since the results of empirical validation ultimately depend on the users' demands, it is recommended that when a user gives the validation conclusions of empirical validation, he/she should provide the 'limit of agreement' simultaneously to guide other users making their own decision.

5.3 Comparing the prediction accuracy of models

The new method for comparing the accuracy of HBT models attempts to improve the traditional RMSE method by applying RMSE based on validation results of the population-based models. In the case study, the conclusion that Model B is more accurate than Model A is drawn because the accuracy of Model B is validated as level Ⅰ while Model A is level Ⅲ. However, if the judgement is purely based on the traditional RMSE Method, the fact that the RMSE of Model A is smaller than that of Model B will lead to a conclusion opposite to the one obtained from the new method. Obviously, a model which is able to make statistically accurate predictions should be

superior to a model which is inaccurate. This result reveals the limitations of the traditional RMSE Method.

Compared with the traditional RMSE Method, we believe that the new method is more general and rational. Comparisons between models are not only based on the comparisons of the RMSEs between predictions and sample means, but also related to the other statistics such as sample standard deviation and sample size. For example, for the case in Section 3, if the sample mean and model prediction remain unchanged, and the standard deviation of each sample widens to 1.5 times as much as before, the confidence intervals for each population mean will be expanded and data in this modified case is shown in Figure 6.

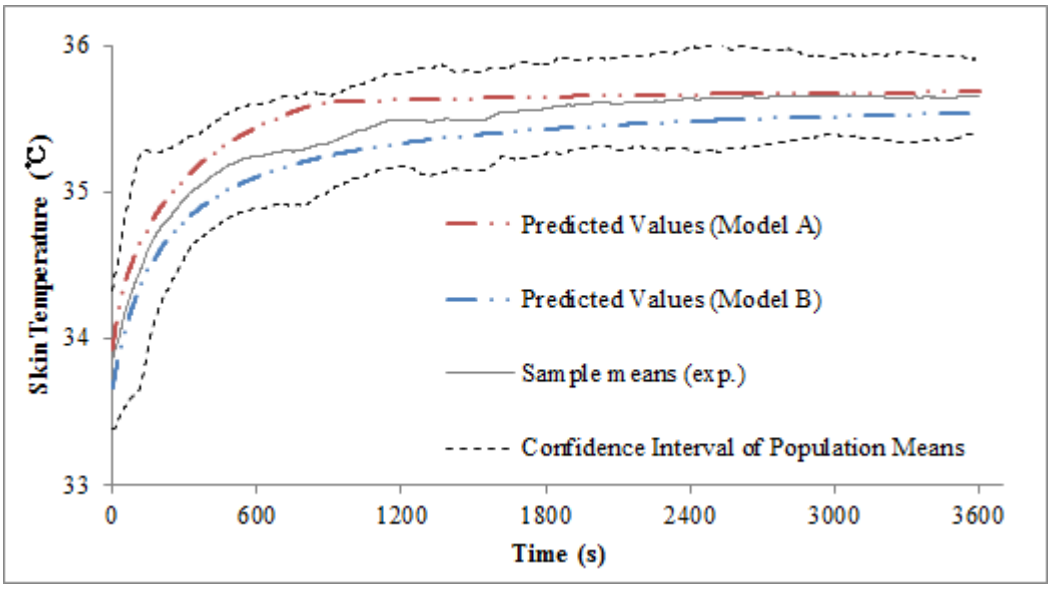


Figure 6. The predictions and statistical information from the samples in the modified case

In this modified case study, the RMSE of Model A and Model B are still 0.16 and 0.165 respectively, just the same as the RMSE values in the original case study. But the models' predictions are all within the confidence interval at this time, which means that both Model A and Model B are statistically accurate and should be classified into accuracy level Ⅰ. By comparing RMSE values from the two models, it will be

concluded that Model A is better than Model B. The modified case study here and the original case study in section 3 have no differences on model predictions and sample means, but the difference in sample standard deviation leads to the opposite conclusion when comparing the models.

5.4 Significance level α

According to the elaboration of the new method, the significance level α determines the 'confidence interval of population means' and 'limit of agreement', so different α values may lead to different evaluation results. In this paper, α is set to 0.05, which is a customary choice in statistics. Apparently, other values such as 0.01 or 0.1 can also be selected, but it must be ensured that α is kept as a consistent figure during the whole evaluation process. Using the same α is a precondition for applying this evaluation method to compare the prediction accuracy of different models.

6. Conclusion

This research proposes a method for evaluating the accuracy of the population-based Human Body Thermoregulation Models. Based on the theory of statistical inference, agreement analysis and error analysis, two key questions in model evaluation study namely: i) 'How to validate the prediction accuracy of models?' ii) 'How to compare the prediction accuracy of models' can be properly solved by using this new method. A framework of the validation process for HBT models has been proposed, which validates a model's accuracy both from the statistical and empirical aspects. Five steps are proposed in the framework in the new method as: i) Define the null hypothesis and alternative hypothesis; ii) Define the confidence intervals for population means; iii) Validate the model's accuracy by comparing model predictions with the confidence intervals of the population means; iv) Validate the model's accuracy by analysing the agreement between the model prediction and the sample mean; v) Compare the accuracy of models based on accuracy level and RMSE. For validation of HBT models,

three levels of accuracy are proposed as: Ⅰ—statistically accurate; Ⅱ—empirically accurate; Ⅲ—inaccurate. This method can promote the development and evaluation of the HBT models, which is very important in the studies of human physiology or thermal comfort. Furthermore, the new method is not only suitable for the evaluation of HBT models, but can also be theoretically applied to the evaluation of population-based models in other research fields.

ACKNOWLEDGEMENT

The authors would like to thank the Major State Basic Research Development Program of China (Program 973) (Project No. 2012CB720100); the National Natural Science Foundation of China (Project No. 50838009); the 111 Project (No. B13041) for their financial support for the research. Yu Yang would like to thank the China Scholarship Council for the sponsorship for a one-year academic visiting study at the University of Reading during 2013-2014.

References

[1] Cheng Y., Niu J., Gao N. Thermal comfort models: A review and numerical investigation. Building and Environment. 2012;47:13-22.
[2] Gagge**,** A.P., Stolwijk**,** J.A.J**.,** Nishi, Ysaunobu. An effective temperature scale based on a single model of human physiological temperature response. ASHRAE Transactions. 1971;77:247-62.
[3] Stolwijk J. A mathematical model of physiological temperature regulation in man: National Aeronautics and Space Administration; 1971.
[4] Tanabe S-i., Kobayashi K., Nakano J., Ozeki Y., Konishi M. Evaluation of thermal comfort using combined multi-node thermoregulation (65MN) and radiation models and computational fluid dynamics (CFD). Energy and Buildings. 2002;34:637-46.
[5] Huizenga C., Hui Z., Arens E. A model of human physiology and comfort for assessing complex thermal environments. Building and Environment. 2001;36:691-9.
[6] Zolfaghari A., Maerefat M. A new simplified thermoregulatory bioheat model for evaluating thermal response of the human body to transient environments. Building and Environment. 2010;45:2068-76.

[7] Munir A., Takada S., Matsushita T. Re-evaluation of Stolwijk's 25-node human thermal model under thermal-transient conditions: Prediction of skin temperature in low-activity conditions. Building and Environment. 2009;44:1777-87.
[8] Fiala D., Lomas K.J., Stohrer M. Computer prediction of human thermoregulatory and temperature responses to a wide range of environmental conditions. International Journal of Biometeorology. 2001;45:143-59.
[9] Xu X., Werner J. A dynamic model of the human/clothing/environment-system. Applied human science: journal of physiological anthropology. 1997;16:61-75.
[10] Gordon R.G., Roemer R.B., Horvath S.M. A mathematical model of the human temperature regulatory system-transient cold exposure response. Biomedical Engineering, IEEE Transactions on. 1976:434-44.
[11] Werner J. Control aspects of human temperature regulation. Automatica. 1981;17:351-62.
[12] Smith P., Twizell E. A transient model of thermoregulation in a clothed human. Applied Mathematical Modelling. 1984;8:211-6.
[13] Haslam R., Parsons K. An evaluation of computer-based models that predict human responses to the thermal environment. ASHRAE Trans. 1988;94:1342–60.
[14] Höppe P.R. Heat balance modelling. Experientia. 1993;49:741-6.
[15] Werner J, Webb P. A six-cylinder model of human thermoregulation for general use on personal computers. Annals of physiological anthropology. 1993;12:123-.
[16] Yi L., Fengzhi L., Yingxi L., Zhongxuan L. An integrated model for simulating interactive thermal processes in human–clothing system. Journal of Thermal Biology. 2004;29:567-75.
[17] Kaynakli O., Kilic M. Investigation of indoor thermal comfort under transient conditions. Building and Environment. 2005;40:165-74.
[18] Salloum M., Ghaddar N., Ghali K. A new transient bioheat model of the human body and its integration to clothing models. International Journal of Thermal Sciences. 2007;46:371-84.
[19] Al-Othmani M., Ghaddar N., Ghali K. A multi-segmented human bioheat model for transient and asymmetric radiative environments. International Journal of Heat and Mass Transfer. 2008;51:5522-33.
[20] Yildirim E.D., Ozerdem B. A numerical simulation study for the human passive thermal system. Energy and Buildings. 2008;40:1117-23.
[21] Wan X., Fan J. A transient thermal model of the human body–clothing–environment system. Journal of Thermal Biology. 2008;33:87-97.
[22] Ferreira M., Yanagihara J. A transient three-dimensional heat transfer model of the human body. International Communications in Heat and Mass Transfer. 2009;36:718-24.
[23] Schellen L, Loomans M., Kingma B., de Wit M., Frijns A., van Marken Lichtenbelt W. The use of a thermophysiological model in the built environment to predict thermal sensation: Coupling with the indoor environment and thermal sensation. Building and Environment. 2013;59:10-22.
[24] Armstrong J.S., Collopy F. Error measures for generalizing about forecasting methods: Empirical comparisons. International journal of forecasting. 1992;8:69-80.

[25] Altman D.G., Bland J.M. Measurement in medicine: the analysis of method comparison studies. Statistician. 1983;32:307-17.
[26] Bland J.M., Altman D.G. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet. 1986;327:307-10.
[27] Bland J.M., Altman D.G.. Measuring agreement in method comparison studies. Statistical methods in medical research. 1999;8:135-60.
[28] Zaki R., Bulgiba A., Ismail R., Ismail N.A. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PloS one. 2012;7:e37908.
[29] Schervish M.J. Theory of statistics: Springer; 1995.
[30] Taki A., Ealiwa M., Howarth A., Seden M. Assessing thermal comfort in Ghadames, Libya: Application of the adaptive model. Building Services Engineering Research and Technology. 1999;20:205-10.
[31] Hyndman R.J., Koehler A.B. Another look at measures of forecast accuracy. International journal of forecasting. 2006;22:679-88.
[32] M. Humphreys, J.F. Nicol, The validity of ISO-PMV for predicting comfort votes in every-day thermal environments, Energy and Buildings, 34 (2002) 667-684.