Individualized Short-Term Core Temperature Prediction in Humans Using Biomathematical Models

Andrei V. Gribok, Mark J. Buller, and Jaques Reifman*

***Abstract*—This study compares and contrasts the ability of three different mathematical modeling techniques to predict individual-specific body core temperature variations during physical activity. The techniques include a first-principles, physiology-based (SCENARIO) model, a purely data-driven model, and a hybrid model that combines first-principles and data-driven components to provide an early, short-term (20–30 min ahead) warning of an impending heat injury. Their performance is investigated using two distinct datasets, a Field study and a Laboratory study. The results indicate that, for up to a 30 min prediction horizon, the purely data-driven model is the most accurate technique, followed by the hybrid. For this prediction horizon, the first-principles SCENARIO model produces root mean square prediction errors that are twice as large as those obtained with the other two techniques. Another important finding is that, if properly regularized and developed with representative data, data-driven and hybrid models can be made "portable" from individual to individual and across studies, thus significantly reducing the need for collecting developmental data and constructing and tuning individual-specific models.**

***Index Terms*—Core temperature prediction, data-driven model, first-principles model, heat injury, hybrid model, regularization, time-series analysis.**

I. Introduction

HEAT injury is the third leading cause of death of student athletes at U.S. schools [1]. Heat injury is also a problem for the armed forces, especially during deployments to localities with very hot climates. Despite thorough prevention programs developed by the U.S. Army Research Institute of Environmental Medicine (USARIEM), from 2003 through 2005, there were over 4401 heat injuries in the armed forces, of which 784 were heat strokes and 3617 were heat exhaustions [2]. There were an additional 17 heat-related fatalities during this time period. Although heat injuries are considered to be preventable, a previously published study showed that humans lack warning mechanisms to signal an impending serious heat injury [3]; hence, in certain situations, a reliable system for real-time continuous monitoring and prediction of body core temperature would be highly desirable. Such a prediction system, coupled with the known clinical limit of 40 °C [4], could potentially prevent heat-related injuries.

Recent advances in the ability to monitor physiology variables have resulted from the development of new biosensors and information-processing capabilities. These capabilities have a direct impact on how closely a person's state can be monitored during civilian activities or during military operations, including the possibility of predicting changes in many vital physiological variables, such as body core temperature, heart and respiratory rates, and even such subtleties as level of alertness and performance. The technological breakthroughs in the development of hardware and firmware were also accompanied by an equally profound and significant progress in such fields as data mining and machine learning. New technologies to collect and store relatively large amounts of physiological data in the field allow researchers to explore new opportunities in data-driven methods to forecast physiological variables and status.

For example, the Warfighter Physiological Status Monitoring (WPSM) program at the U.S. Army Medical Research and Materiel Command seeks to develop a soldier-wearable, computer-based system for providing commanders and medics with critical physiological status information about dismounted war fighters [5], [6]. The WPSM system has two primary aims: the first is to prevent nonbattle injuries, such as heat stroke and dehydration, and the second is to optimize casualty management through improved casualty detection, diagnostics, and triage. These aims require an array of sensors, a personal area network, and data management software as well as a variety of decision-support algorithms for monitoring and predicting a soldier's physiological status. In this paper, we focus on mathematical modeling techniques that can be used to prevent impending nonbattle heat injuries, such as heat exhaustion and heat stroke. We compare and contrast the ability of three types of models (a first-principles model, a purely data-driven model, and a hybrid model that combines first-principles and data-driven components) to produce short-term (20–30 min), individual-specific

Manuscript received March 27, 2007; revised September 11, 2007. This work was supported in part by the Combat Casualty Care and the Military Operational Medicine Research Programs of the U.S. Army Medical Research and Materiel Command, Fort Detrick, MD and in part by the Natick Soldier System Center. *Asterisk indicates the corresponding author.*

A. V. Gribok is with the Bioinformatics Cell, Telemedicine and Advanced Technology Research Center (TATRC), U.S. Army Medical Research and Materiel Command (USAMRMC), Fort Detrick, MD 21702 USA. He is also with the Nuclear Engineering Department, University of Tennessee, Knoxville, TN 37996 USA (e-mail: agribok@bioanalysis.org).

M. J. Buller is with the U.S. Army Research Institute of Environmental Medicine (USARIEM), Natick, MA 01760 USA. He is also with Brown University, Providence, RI 02912 USA (e-mail: mark.j.buller@us.army.mil).

*J. Reifman is with the Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center (TATRC), U.S. Army Medical Research and Materiel Command (USAMRMC), Fort Detrick, MD 21702 USA (e-mail: jaques.reifman@us.army.mil).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2007.913990

predictions of body core temperature variations during physical activity.[1]

Physiological models commonly rely on first-principles knowledge about various mechanisms in the human body and their associated dynamics. Although some underlying physiological phenomena are not well understood and are therefore unmodeled, the resulting first-principles models may still be effective in predicting some population-average responses with certain fidelity. However, unless the model parameters are constantly adjusted, based, for example, on measurements from a specific individual, in general, first-principles models are not capable of representing interindividual variability [7], [8], leading to inaccurate predictions for specific individuals. Individuals with similar anthropomorphic characteristics and subject to the same workload and environmental conditions may yield very different physiological responses. Interindividual variation in physiological response is particularly critical at limiting thresholds of physiological health, such as at extreme values of core temperature, where small variations can make a difference between a suitable recovery and an irreversible pathological condition. The need to represent interindividual variability can be addressed by developing models that utilize historic and current data that are specific to the individual.

One approach to improve the fidelity of first-principles models and account for interindividual variability is to incorporate data-driven or "black box" models into the first-principles model to create a "hybrid" model [9]. In this case, the data-driven portion of the hybrid model is intended to capture the dynamics and the physiological idiosyncrasies of each particular individual, which the first-principles model cannot capture, by "learning," during the "training" phase, the residuals between predictions produced by the first-principles model and the actual measurements. This allows hybrid models to account for interindividual variability and also for parts of the poorly modeled dynamics. The hybrid approach was introduced to the physiological community in a previous study [9], where different hybrid schemes were presented and contrasted. Hybrid models have been widely used in system identification and control in industrial processes and have proven to be quite effective [10], [11]. Hybrid modeling of physiological dynamics holds equal promise in this regard [12], [13].

Another approach to physiological predictions is to employ a purely data-driven model. A stand-alone, data-driven model can be trained on historical data, and subsequently used to predict future unknown data. The historical data can include independent variables related to the predicted variable as well as delayed instances of the predicted variable itself, that is, previous core temperature measurements in this case. An inherent limitation of purely data-driven models is their inability to extrapolate reliably beyond the distribution of the "training" data. However, linear data-driven models are quite often good extrapolators if the underlying dependencies can be reasonably modeled by linear laws. Furthermore, many physiological variables are tightly bounded by homeostatic limits, thus simplifying the problem of collecting data that cover all physiologically plausible situations. These provide an opportunity to properly train linear data-driven models on representative samples of historical data and determine their generalization effectiveness, including their ability to be made "portable" from one individual to another.

Another general limitation of data-driven modeling is the possibility of "excessive explanation" of the training data, leading to an "overfitted" model with poor generalization capabilities. The problem of overfitting is quite often understated in the case of linear data-driven models; however, this effect is as detrimental in linear models as it is in their nonlinear counterparts. This paper demonstrates that proper regularization of purely data-driven models and the data-driven portion of hybrid models is crucial to their generalization capabilities, since it precludes overfitting and produces models that capture the underlying data dependencies but not their idiosyncrasies.

II. Methods

A. First-Principles SCENARIO Model

The first-principles SCENARIO model [14], [15], developed at USARIEM, was designed to estimate and predict core temperature, heart rate, and sweat rate, without requiring prior knowledge and direct measurement of these physiological variables. The underlying model for SCENARIO simulates the time course of core temperature variations, while taking into account different factors that affect human thermoregulation. The temperature distribution within the human body is represented by a lump-parameter model consisting of six concentric cylindrical compartments. Heat flow is then modeled by a set of macroscopic energy conservation equations based on heat convection between the central blood compartment and the adjacent core, muscle, fat, and vascular skin compartments; radial heat conduction between every pair of adjacent compartments; and air convection, radiation, and sweat evaporation between the superficial avascular skin layer and the environment and transition through the clothing [14], [15]. The energy conservation equations are represented by a set of six ordinary differential equations that can be expressed as

$$\frac{dT}{dt} = A(t)T(t) + B(t) \tag{1}$$

where $T(t) \in R^{6\times 1}$ is a vector representing the bulk temperatures in each of the six modeled compartments, and $A(t) \in R^{6\times 6}$ is a time-varying matrix determined by parameters, such as the conductance between two adjacent compartments and blood flow between the compartments. The vector $B(t) \in R^{6\times 1}$ accounts for the secondary inputs to the system, and it is primarily governed by the metabolic rate in each of the compartments, as well as the respiration rate. The various

[1] In collecting the data presented in this manuscript, the investigators adhered to the policies for protection of human subjects as prescribed in Army Regulation 70–25, and the research was conducted in adherence with the provisions of 45 CFR Part 46. The subjects gave their informed consent for the laboratory study after being informed of the purpose, risks, and benefits of the study.

factors that affect human thermoregulation and used as input to SCENARIO include:

1) environmental: mean radiant temperature, ambient temperature, relative humidity, wind speed;
2) activity: walking speed, pack weight (load), terrain factor, slope/grade, water intake;
3) individual characteristics: age, weight, height, fat percentage;
4) clothing: insulation and permeability.

Being a first-principles model, SCENARIO does not use past temperature measurements to produce future core temperature predictions. Another advantage is that, based on the range of applicability of each underlying model component, the range of applicability of the overarching model can be determined *a priori*. In addition, SCENARIO can predict other physiological variables, such as heart and sweat rates. However, because SCENARIO was designed as a mission-planning tool, as opposed to an early thermal warning system, it is not expected to perform as well as customized models for short-term temperature predictions, where core temperatures are highly correlated. Although SCENARIO's input parameters are specific to an individual's characteristics, internally, it does not represent parameter model differences to fully account for interindividual variability. Additionally, since all parameters are estimated on the basis of experimental data, inherent observation error and limited sample size may lead to discrepancies that, compounded, could contribute to model inaccuracy. Furthermore, due to simplifying modeling assumptions and unmodeled (unknown) physiology, SCENARIO does not fully represent some of the physiological dynamics. Hence, SCENARIO is partly used here as a benchmark, and it is selected among other first-principles models [16]–[18] because it has been traditionally used by the Army to analyze the human response to heat stress and was readily available to the authors. We acknowledge, however, that the reported results are only applicable to SCENARIO and cannot be generalized to other first-principles models, which may demonstrate better performance under similar conditions.

B. Data-Driven Modeling

Data-driven linear models have been used for time-series prediction since the early 1970s [19]. One of the most widely used linear models is the autoregressive (AR) model [10], which allows for the inference of estimates $\hat{y}_n$, at time *n*, $n = m+1, \ldots, N$, of signal *y* as a function of previous observations

$$\hat{y}_n = \sum_{i=1}^{m} b_i y_{n-i} + \varepsilon_n. \tag{2}$$

where *b* represents the vector of AR coefficients to be determined, ε_n denotes white noise with unknown variance, *N* denotes the number of data samples, and *m* is the order of the model, i.e., the number of previous measurements used to predict the future measurement $\hat{y}_n$. Interchanging $\hat{y}_n$ for y_n, and defining the $(N-m) \times (m)$ design matrix *U* and the

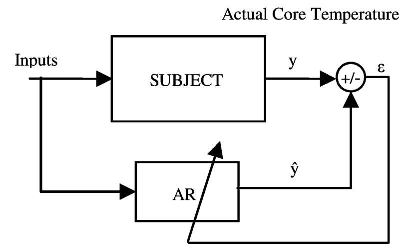

Fig. 1. Data-driven approach to physiological time-series prediction; *y* is the actual core temperature measurement, $\hat{y}$ is the predicted core temperature, ε is the residual between the measured core temperature and the predicted core temperature, and inputs represent exogenous data into to the system, such as ambient temperature and past measurements of core temperature. The crossing of the AR box signifies that the AR coefficients are computed during the training phase.

$(N-m) \times (1)$ and $(m) \times (1)$ vectors *y* and *b*, respectively, as

$$U = \begin{bmatrix} y_m & y_{m-1} & \cdots & y_1 \\ y_{m+1} & y_m & \cdots & y_2 \\ \vdots & \vdots & \ddots & \vdots \\ y_{N-1} & y_{N-2} & \cdots & y_{N-m} \end{bmatrix},$$

$$y = \begin{bmatrix} y_{m+1} \\ y_{m+2} \\ \vdots \\ y_N \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{3}$$

we arrive at an overdetermined system of linear equations. This system can be solved for *b* by the least-squares (LS) method, which seeks *b* that minimizes

$$\underset{b}{\operatorname{argmin}} \, \|y - Ub\|^2 \tag{4}$$

provided the design matrix *U* is well-conditioned. In addition to the estimation of the coefficients *b*, the model's order also needs to be determined, which can be done by using some analytical criterion, like the minimum description length approach [20] and Akaike information criterion [21], or by cross-validation.

Data-driven models are generally used in problems where obtaining a first-principles model is either impractical or difficult due to excessive complexity of the underlying phenomena to be modeled, and it was a motivating factor for this study. A schematic diagram of the data-driven approach is presented in Fig. 1. The advantage of the data-driven approach is that the explicit relationships between the input–output variables in the modeled phenomenon do not need to be known and can be "learned" during the "training" phase. The approach, however, is highly dependent on data availability and on the quality of the available data. Another difficulty is that learning input–output dependencies from noisy data constitutes an ill-posed problem, since several models may explain the training data quite well, generally due to model overfitting, although not all models will posses good generalization capability.

Data-driven models can also be nonlinear, represented by artificial neural networks (ANNs), for example. The difference between AR and ANN models is that AR models can only capture linear dependencies present in the data, while ANNs can also accommodate nonlinear relationships. However, due

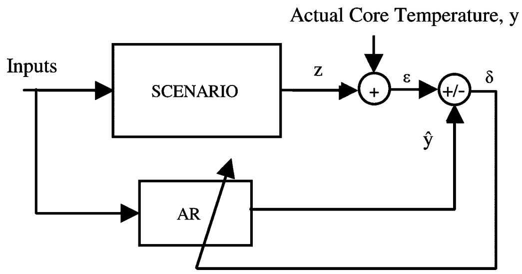

Fig. 2. Hybrid approach to physiological time-series predictions; z is the SCENARIO core temperature estimates, y is the actual core temperature measurements, $\hat{y}$ is the predicted value of the residual ε (i.e., the difference between the SCENARIO prediction and the core temperature measurement), inputs are exogenous inputs to the system, and δ is the residual between $\hat{y}$ and ε.

to the presence of local minima in the cost function, ANNs may be harder to train. Also, in cases where the process input–output dependencies are linear, they provide no added benefit. In a previous core temperature prediction study by our group, ANNs failed to outperform linear models [13].

C. Hybrid Modeling

Another modeling approach is the hybrid technique, which tries to capitalize on the best parts of both models—first-principles and data-driven. The general idea of hybrid modeling is presented in Fig. 2, where the data-driven component is represented by an AR model. The hybrid approach first attempts to predict a data value for a physiological variable using the first-principles model. The residual value ε of this prediction (i.e., the difference between the first-principles model prediction and the measured value) is then presented to a data-driven model as a target signal, and the data-driven model is trained to fit ε based on its past values and, possibly, exogenous inputs. After the training is complete, new data are predicted by adding the predictions of the first-principles model with those of the data-driven model. In our implementation, only delayed instances of the residual signal ε are used as inputs to the data-driven portion of the hybrid.

An important difference from the purely data-driven approach is that, in hybrid modeling, the data-driven component learns the residuals between the first-principles predictions and the actual measurements, while in purely data-driven modeling, the model learns the actual measurements. Several arguments have been put forward to justify the use of hybrid modeling for time-series predictions. For example, it was shown previously [22] that, provided the first-principles model has the same form as the true process model, the hybrid is guaranteed to converge to the true model as the amount of training data increases indefinitely. Another argument is that the residuals may be easier to learn than the actual measurements [23] because the residuals only cover a subspace of the whole process space. Significant successes in applying hybrid models have been reported in chemical and biochemical engineering [23]. However, to produce accurate predictions, hybrid models require high-fidelity first-principles models capable of accurately predicting both the training data and the testing data. If they fail to produce good predictions for the training data, the target signal for the data-driven part of the hybrid will not be adequate. Also, if they fail to accurately model the testing data, the hybrid predictions will not be accurate, since in this case, the overall prediction error will be dominated by the error produced by the first-principles component of the model.

D. Regularization of the Data-Driven and Hybrid Models

As mentioned earlier, fitting a data-driven AR model to data (either as a stand-alone module or as part of a hybrid model) requires estimation of the AR coefficients as one of the steps. The coefficients are usually determined by minimizing the LS functional in (4). Unfortunately, due to the highly correlated nature of the core temperature signal, the design matrix U is quite often ill-conditioned or even numerically rank deficient. This causes the estimates of the AR coefficients b to be highly unstable, producing poor-quality predictions, i.e., degraded generalization. The reason for the degraded performance is that the unconstrained minimization of (4), when U is ill-conditioned, causes the solution to be dominated by high-frequency components that overfit the training data [24]. The practical consequences of the ill-conditioning of the design matrix U are demonstrated in Section III.

It is well known that the LS solution to (4) yields an unbiased estimator with the smallest variance among unbiased estimators [25]. Although unbiasedness is intuitively desired, in practice, it could be quite useless due to the potential large variance of the unbiased estimator. To deal with this problem, a class of biased estimators known as regularized least squares was proposed by Tikhonov [24]. In this method, the minimization of (4) is replaced by the minimization of the augmented functional

$$\underset{b}{\operatorname{argmin}} \|y - Ub\|^2 + \lambda^2 \|Lb\|^2 \tag{5}$$

where the regularization parameter λ controls the tradeoff between the smoothness of the solution and its fit to the training data, and L is a well-conditioned matrix; for example, a discrete approximation of a second-order derivative operator was used in this study. The major benefit of the regularized LS estimate is that it reduces the variance of the solution by introducing a small bias to generate a much smaller estimation error, defined as the variance plus the square of the bias between the true (unknown) parameter and its estimate [26].

E. Datasets

We employed two datasets to develop, compare, and contrast the three modeling approaches: Field (dataset A) and Laboratory (dataset B).

1) Field Study (Dataset A): The Field dataset [15] consists of physiological data collected from eight U.S. Marine Corporations volunteers [age: 25 year (SD 3.2); height: 174 cm (SD 6.7); weight: 71.6 kg (SD 7.9); body fat pct: 15.9% (SD 7.1), mean and standard deviation (SD)] during a four-day field exercise. Each 10 h day involved a 3 mi morning march to a shooting range, followed by day-long exercises and rotations within firing stations, and a march back via the same route in the evening. Subjects wore air-permeable battle dress uniform (thermal resistance $= 1.32\ \mathrm{m}^2\cdot\mathrm{K/W}$) and, when marching, carried a pack load of 26±1.0 kg. The ground temperature during the day was 29.8 °C (SD 0.5), and the dew point and wind speed were 21.1 °C (SD 0.5) and 4.2 m/s (SD 0.5), respectively. The core temperature for each subject was measured through a telemetry pill ingested at the beginning of each day. There is a close relationship among core temperatures measured by esophageal probes, rectal probes, and telemetry pills during exercise activities in both temperate and hot conditions [27].

Unfortunately, sometimes the signal from the pill could not be detected, and other times, the pill produced very noisy temperature signals. To eliminate data artifacts and reduce noise levels, the temperature data are preprocessed using median and moving-average filters. The median filter is used for its known outlier rejection capabilities, and the smoothing filter is used to remove high-frequency signal noise and to interpolate short regions of missing values. The core temperature was recorded every minute for each of the 10 h days for each of the four days.

2) Laboratory Study (Dataset B): The Laboratory-based dataset [28] consists of core temperature measurements collected from nine volunteer subjects [age: 23 year (SD 4); height: 174.2 cm (SD 5.8); weight: 73.4 kg (SD 6.5); body fat pct: 17.9% (SD 3.99)], whose anthropomorphic characteristics are very similar to those of the Field study, dataset A. The subjects walked on a treadmill under two environmental conditions: control (day 1: 20 °C temperature and 50% relative humidity) and humid (day 2: 27 °C temperature and 75% relative humidity). The wind speed was 1.1 m/s for both conditions. On the morning of the test days, the subjects, dressed in air-permeable battle dress uniform with the same thermal resistance as in the field study, were instrumented for the collection of various physiological variables, including core (rectal) temperature. Next, they sat on a chair for 10 min just before starting to walk at 3 mi/h on level treadmills. The walking paused after every 30 min for 10 min of sitting. There were four 30 min walking periods/test so that the entire experiment lasted a total of 170 min, including 10 min rest periods at each end. At the end of each 10 min pause, the subjects were given 150 mL of water before walking again. Rectal temperature (assumed to be representative of the core temperature) was collected continuously and recorded every minute, as in dataset A.

Typical temperature measurements for two subjects for each of the two datasets are presented in Fig. 3. Notice that the standard deviation of the core temperature signal in the Field study, dataset A, is two times larger than that of the Laboratory study, dataset B. Dataset A also has a larger amount of data, which is reflected by the different scales on the time axes in Fig. 3.

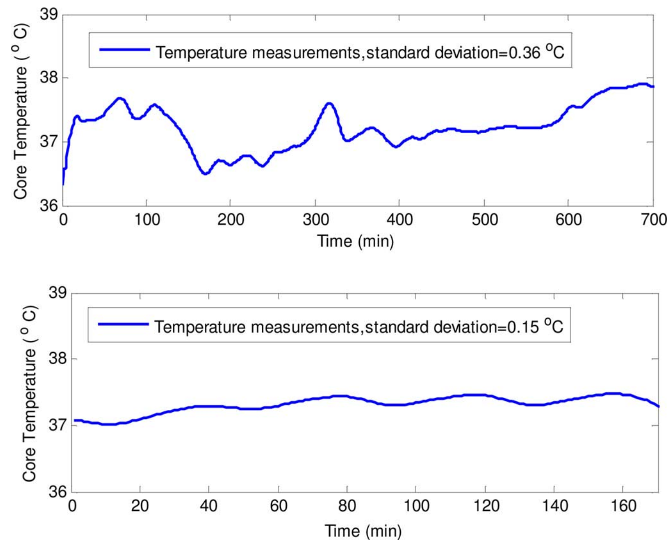

Fig. 3. Temperature profiles for two subjects from the two datasets used in the simulations. Top: Field study, dataset A, Bottom: Laboratory study, dataset B. Note different scales on the *x*-axis.

F. Simulation Tests

Four different computer simulations are considered, which are referred to as simulations S1, S2, S3, and S4.

1) S1: Same-subject simulation. For each of the eight subjects in dataset A, a data-driven model and a hybrid model are separately trained on one (randomly selected day) of the four days of each subject's data, resulting in 16 (8×2) different models. The subject-specific models so developed are tested on that subject using the remaining three days of available data. The SCENARIO model is separately applied to the corresponding three days, as in the testing of the data-driven and hybrid models, for each of the eight subjects.
2) S2: Cross-subject simulation. Sixteen models are developed as in simulation S1 earlier, and then, the models are tested on all four days of the other seven subjects' data. That is, each model is blind to the subject's data it is tested on. The SCENARIO model is separately applied to all four days of each of the seven subjects used for testing.
3) S3: Cross-study A–B simulation (train on dataset A and test on dataset B). Sixteen models are developed as in simulation S1 earlier, and then, the models are tested for both days of each of the nine subjects in dataset B, that is, each model is blind not only to the subject it is tested on but also to the study itself. The corresponding SCENARIO simulations are run separately for each of the two days for each of the nine subjects in dataset B.
4) S4: Cross-study B–A simulation (train on dataset B and test on dataset A). Similar to simulation S3 but 18 instead of 16 models (one data-driven and one hybrid model for each of the nine subjects of dataset B) are developed using data from the first day of the Laboratory study, and subsequently, tested on all four days and eight subjects in dataset A. The corresponding SCENARIO simulations are run separately for each of the four days for each of the eight subjects in dataset A.

For all simulations of the data-driven and hybrid models, the prediction horizon, unless otherwise noted, is set to 20 min. The 20 min ahead prediction horizon is selected based on its practical utility, since it provides sufficient warning time to prevent thermal stress injuries while allowing the models to produce data-driven predictions of acceptable accuracy. Note that there

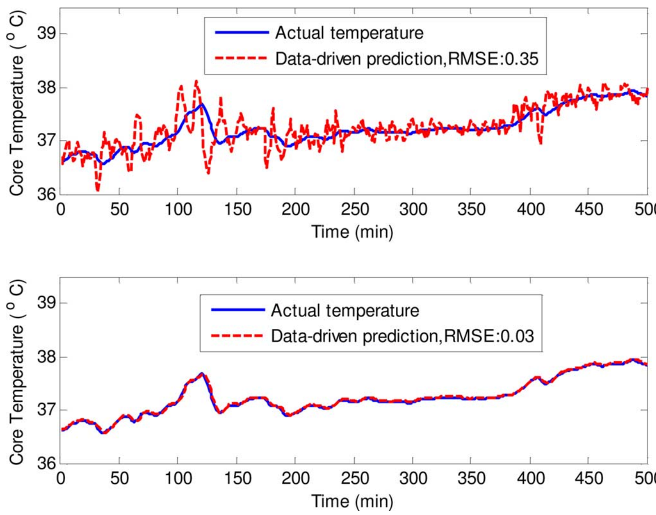

Fig. 4. Top: unregularized core temperature predictions using AR models, bottom: regularized AR models. The solid curve represents the actual measured core temperature for day 1, subject #1 in dataset A. The dashed curves represent 20 min ahead temperature predictions obtained with unregularized and regularized AR models of order 25, trained on day 2 data from subject #1.

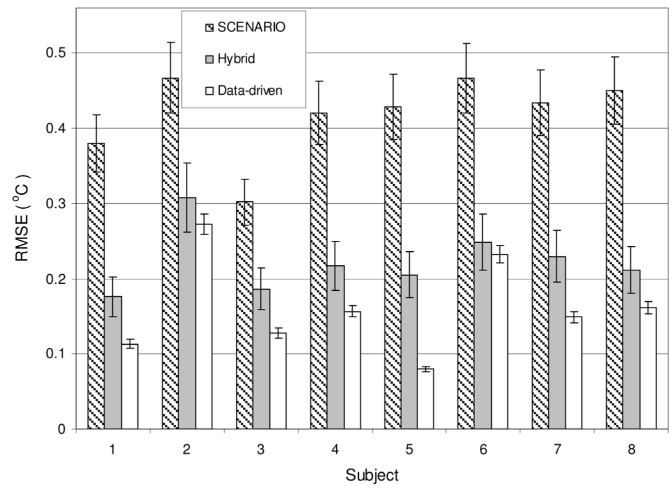

Fig. 5. RMSE for the same-subject (simulation S1) predictions. For the hybrid and data-driven models, each bar represents average RMSE of the model's predictions for that individual over the three days that are not used for training. For SCENARIO, the bars represent the average RMSE for the corresponding three days. The error bounds correspond to one standard deviation. The standard deviation is calculated over all testing subjects and all testing days.

is no prediction horizon to speak of for SCENARIO, since it iteratively computes the entire temperature profile over the desired time length. Each run is performed for a specific individual, i.e., it does not perform cross-individual predictions, since the input data to SCENARIO correspond to the individual's data it is predicting. Also, SCENARIO requires numerous independent variables as inputs, such as walking speed, terrain, slope/grade, and water intake, which we assume to be known.

Through experimentation with different model structures, we determined that a simple AR model suffices to predict core temperature for the purely data-driven models and to predict residuals for the hybrid models. The order of the models is selected using a cross-validation approach, and it is determined that, for all the subjects in dataset A, the overall optimum order is around 25. An AR model of this order is used for both data-driven and hybrid models for both datasets. The adequacy of the models is verified by checking for whiteness of the residuals. We find that the residuals' autocorrelation function consistently lies within the 99% confidence intervals, thus confirming that the models correctly describe the data. The core temperature data (as well as the residuals used in the hybrid model) are also detrended before application of the AR models to ensure stationarity.

The coefficients of the AR models are estimated using the regularization technique described earlier. As pointed out in Section II, unregularized models produce highly inconsistent predictions, as shown in the top graph in Fig. 4, whereas regularized predictions (Fig. 4, bottom) are much smoother and overlap the actual measurements. Fig. 4 shows the AR model's 20 min ahead predictions for day 1 (subject #1 in dataset A), where the models are trained on day 2 data for the same subject.

Notice the oscillatory nature of the unregularized core temperature predictions and the dramatic change in the quality of the predictions after the model is regularized, reflected in a much smaller root mean square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{6}$$

where $\hat{y}$ and y are the predicted and measured core temperature values, respectively, and N is the number of samples. Although this example is for a purely data-driven model, the same holds for the data-driven counterpart of the hybrid model. All of the results presented here are based on regularized models with the regularization parameter λ in (5) selected by employing the discrepancy principle [29].

III. Results and Discussions

Simulation S1 is devised as a basic test to determine whether data-driven and hybrid models trained on portions of the data for a given individual are able to predict other portions of the same individual's data not used for training. To accomplish this, for each of the eight subjects in dataset A, we develop one data-driven model and one hybrid model using data from one (randomly selected) day out of the four days of the Field study. Each model is then applied to predict, 20 min ahead, the core temperature for the corresponding individual for the remaining three days. The RMSE for each subject's predictions, for each model, is calculated for all three days and averaged. The SCENARIO's RMSE is calculated as the average over the corresponding three days for each subject. The RMSEs for the three models are presented in Fig. 5 along with the error bounds corresponding to one standard deviation.

These results show that data-driven and hybrid models can generalize well if each model is applied to predict the subject for which it is developed, even if model training and testing are performed to data collected on different days. Although these results are promising, their general applicability would require

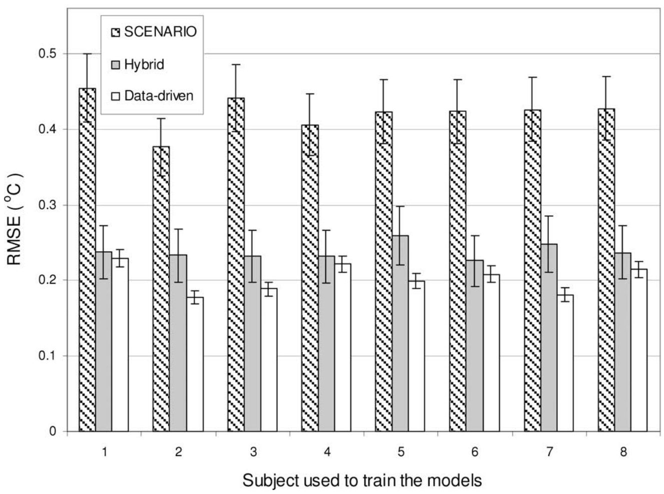

Fig. 6. RMSE for cross-subject (simulation S2) prediction. Each bar represents average prediction errors (RMSEs) over the other seven subjects for the AR models (both data-driven and hybrid) trained using that individual's data. Similarly, the SCENARIO RMSEs represent prediction errors averaged over the prediction of the other seven subjects. The error bounds correspond to one standard deviation. The standard deviation is calculated over all testing subjects and all testing days.

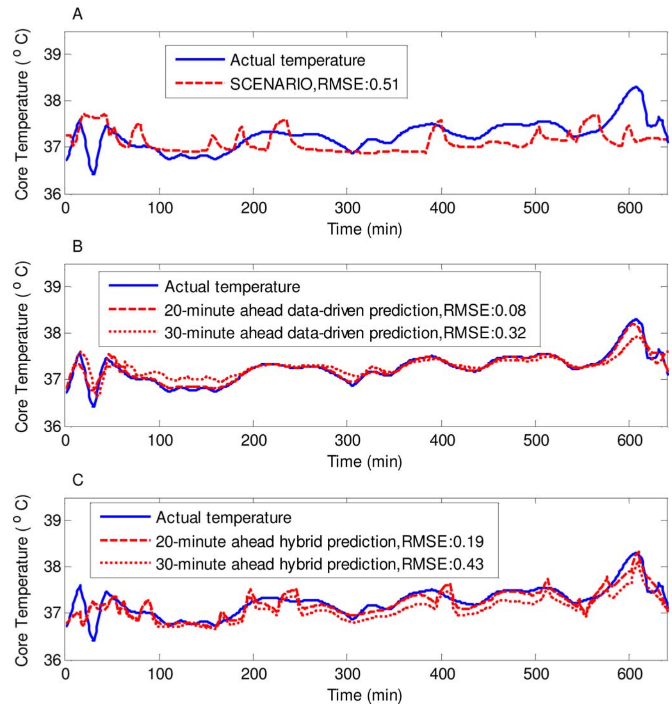

Fig. 7. Core temperature predictions for subject #1, day 2 of dataset A, using three different modeling techniques [(A) SCENARIO, (B) data-driven, and (C) hybrid]. The solid curve represents the measured core temperature. The dashed and dotted curves [(B) and (C)] represent 20 and 30 min ahead predictions, respectively, obtained with the hybrid and data-driven techniques, which are trained using day 1 data from subject #6.

separate collection of core temperature training data for each individual, which is not desirable for practical applications. The most useful application of the data-driven and hybrid techniques comes from the possibility of developing models for one subject and using them to predict different subjects, thus making data-driven models "portable" from one individual to another and reducing the need for data collection.

To test this hypothesis, we perform the cross-subject simulation S2. Fig. 6 illustrates the RMSEs for the three modeling approaches, where for each one of the 24 (8×3) models the RMSEs are averaged over the four days and seven subjects that the models are applied to predict core temperature.

Although the prediction errors for the data-driven and hybrid models are slightly higher than those in Fig. 5, they are still smaller than SCENARIO's RMSEs.

Fig. 7 shows a typical temperature profile prediction for the cross-subject simulation S2.

The predictions are for the second day of subject #1, where the hybrid and the data-driven models are trained with data from the first day of subject #6. Predictions are for both 20 and 30 min horizons. As indicated in the figures and the corresponding RMSEs, the quality of the hybrid and data-driven predictions is highly dependent on the prediction horizon. As expected, the longer the horizon is, the larger is the prediction error. The SCENARIO predictions are obtained by providing input data for subject #1 and having the code consecutively predict the entire temperature profile for that subject at 1 min intervals.

The most challenging set of experiments consists of the two cross-study simulations, S3 and S4, where the data-driven and hybrid models are trained on dataset A and tested on dataset B, and *vice versa*. These two cases are especially challenging because, in addition to using these models to predict "unseen" subjects, the two datasets were collected under very different conditions, where the subjects performed significantly different activities. The results for simulation S3, where dataset A is used to develop the models that are subsequently tested on dataset B, are presented in Fig. 8. The results of S4, where the roles of datasets A and B are reversed, are illustrated in Fig. 9. The results in Fig. 8 indicate that the AR-based models (hybrid and data-driven) are consistently and significantly better than SCENARIO. Interestingly, but perhaps not surprisingly, the predictive performance of the three models is quite different in Fig. 9, where none of the approaches indicates a clear advantage over the others.

The results of this study provide interesting insights into the modeling capabilities of the three different techniques. Specifically, for simulation S1, the average RMSEs (mean and SD) for the three different techniques are: SCENARIO 0.41 °C (SD 0.05), hybrid 0.22 °C (SD 0.04), and data-driven 0.16 °C (SD 0.06). These results suggest that, when large amounts of data from a given subject are available to train data-driven and hybrid models, their predictive capabilities can be quite good. The results also indicate that these models can generalize well across different training and testing days without jeopardizing the models' predictive capabilities. The first-principles SCENARIO model is third in predictive performance, which indicates that, due to the complexity of the thermoregulatory mechanisms in the human body, not all physiological factors can be accounted for in metabolic rate calculations, which are used by SCENARIO as an intermediary step during core temperature

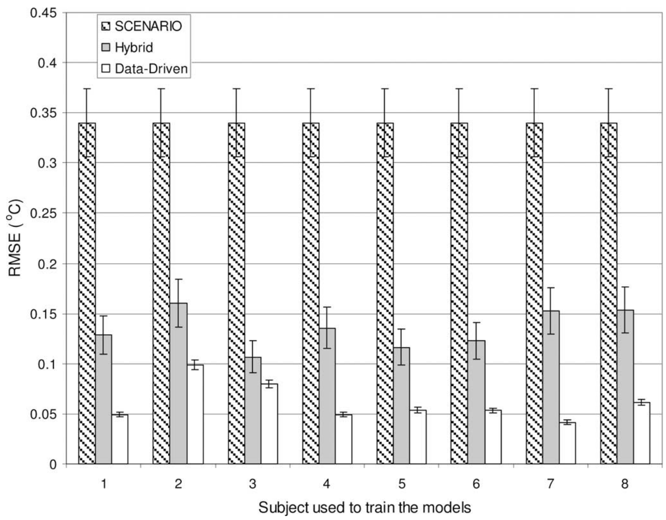

Fig. 8. RMSE for cross-study A–B (simulation S3) predictions. Each bar represents average prediction errors over the nine subjects and two days of dataset B data for AR models (both data-driven and hybrid) trained using the data of the subjects from dataset A indicated on the abscissa. The SCENARIO bars (indicating the same value) represent averaged RMSEs over the two days and all nine subjects in dataset B. The error bounds correspond to one standard deviation. The standard deviation is calculated over all testing subjects and all testing days.

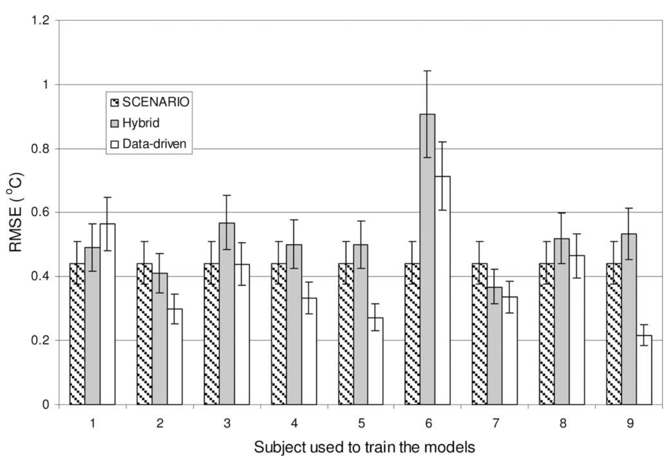

Fig. 9. RMSE for cross-study B–A (simulation S4) predictions. Each bar column represents average prediction errors over the eight subjects and four days in dataset A data for AR models (both data-driven and hybrid) trained using the data of the subjects from dataset B indicated on the abscissa. The SCENARIO bars (indicating the same value) represent averaged RMSEs over the four days and all eight subjects in dataset A. The error bounds correspond to one standard deviation. The standard deviation is calculated over all testing subjects and all testing days.

estimation [30]. Equations modeling sweating, shivering, and vasoconstriction/vasodilatation can also be sources of error.

Simulation S2 is probably a more practically important study, since it examines the situation where a model developed for one individual, using the individual's data, is applied to predict the temperature of other individuals from the same study performing similar activities. For simulation S2, the average RMSEs are: SCENARIO 0.42 °C (SD 0.02), hybrid 0.23 °C (SD 0.01), and data-driven 0.20 °C (SD 0.02). As expected, due to interindividual variability not accounted for in the AR model coefficients, the performance of the data-driven model, in particular, worsened in comparison with simulation S1, but not significantly.

Simulation S3 is quite revealing in terms of the importance for data-driven and hybrid models to have an adequate amount and range of data variability to train the model's coefficients. Because the range of variability of core temperature measurements is much greater in the Field dataset A than in the Laboratory dataset B, as illustrated in Fig. 3, models trained with the former are capable of predicting the latter well. This is reflected by the low average RMSEs for simulation S3 [SCENARIO 0.34 °C (SD 0.03), hybrid 0.13 °C (SD 0.01), and data-driven 0.06 °C (SD 0.01)], which are significantly lower (for the hybrid and the data-driven models) than those in simulations S1 and S2. This suggests that these two modeling approaches may provide more accurate predictions across a different study than that used to develop the models as long as the different study has a narrower range of temperature distribution than the original study. It also suggests that the large amount and range of data variability is able to offset interindividual variability detriments in modeling accuracy.

Simulation S4 illustrates the flip side of this situation. The RMSEs for simulation S4 are as follows: SCENARIO 0.44 °C (SD 0.15), hybrid 0.53 °C (SD 0.15), and data-driven 0.40 °C (SD 0.16). The performance of the data-driven and hybrid models deteriorates when they are trained with laboratory study data, consisting of a limited amount and narrow range of temperature variability, and used to predict core temperature of subjects from the Field study. This simulation clearly reveals two data requirements for the development of "portable" models: 1) availability of large amounts of past temperature measurements and 2) significant range of data variability, encompassing the range of temperatures to be predicted. These are corroborated by recent findings where data-driven models were found to generalize well and be made "portable" when applied to the subjects of the Laboratory study, dataset B [31]. It also suggests, as inferred previously [15], that controlled laboratory datasets may not adequately reflect the true variability of core temperature in the field and should be used with caution when applied to develop models for field use.

The poor performance of the hybrid in simulation S4, due to the limited range and amount of data, is caused by the inability of the data-driven portion of the model to properly learn the residuals during training. The performance of the purely data-driven model deteriorated significantly. However, it is still able to learn the correlations in the training signal and to produce an RMSE that is lower than that of the first-principles SCENARIO model.

The hybrid model displays a middle range performance in terms of prediction accuracy. This can be explained by observing that the residual signal, which is generated by taking the difference between the measured and the SCENARIO-predicted temperature, could be harder to "learn" (by the AR model component of the hybrid) than the temperature measurements. This situation arises when the first-principles component of the hybrid model does not adequately describe the data for a given individual, yielding "random" residual signals that cannot be learned or predicted by the AR portion of the model.

Conversely, when the first-principles model explains the data well, the unexplained portion of the data, i.e., the residuals, will become white noise with no autocorrelation to learn. Obviously, in this case, there is no need to use a hybrid approach. Hence, the hybrid should be used in situations where the first-principles model can successfully explain part of the data but leaves some amount of data unexplained, perhaps the one due to interindividual variability. Also, our results indicate that the hybrid performs well as long as the order of the AR model used to characterize the testing data does not significantly differ from the order of the model needed to characterize the training data.

The power of the purely data-driven approach for near-term predictions comes from the nature of the core temperature signal and the thermal inertia of the human body thermoregulatory process. The low-frequency and smooth nature of the signal lends itself perfectly to AR modeling and predictions, which together with the variability constraints imposed by regularization, force the model to produce core temperature outputs with low variation and excellent predictive capabilities. The relatively large inertia (or time constant) of the body thermoregulatory process is what allows the AR model to make accurate predictions minutes ahead. The thermal inertia, characterized by the specific heat capacity of the human body, regulates and precludes rapid changes in core temperature. This can be explained, for example, by noting that a significant percentage of the human body (up to 75%) is composed of water and that water has one of the largest specific heat capacities of all substances. This large specific heat capacity allows the human body to absorb a significant amount of energy before its temperature rises, thus permitting accurate short-term predictions.

Data-driven models rely entirely on the autocorrelations of the core temperature signal, which do not to exhibit large interindividual variability in our studies, provided individuals are involved in similar activities. As illustrated in Fig. 7, the model accuracy deteriorates as the prediction horizon increases and extends beyond the time constant of the thermal inertia of the human body thermoregulatory process, estimated by us to be around 15 min. This fact is demonstrated in Fig. 10, where a typical autocorrelation function of the actual temperature measurements and the calculated SCENARIO residuals are plotted as a function of time lag.

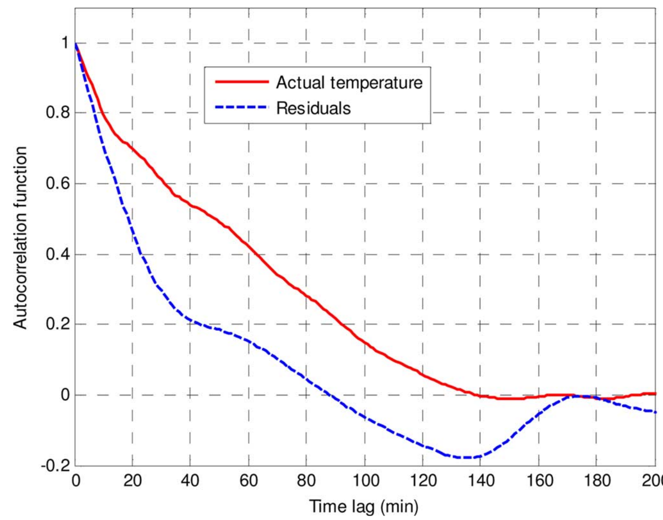

Fig. 10. Autocorrelation functions of temperature measurements and residuals between SCENARIO estimates and actual temperature measurements. The function shows how consecutive data points of the time series are correlated with each other as a function of the distance (time delay) in the time series.

The autocorrelation function shows how quickly the correlation between samples decays as a function of time and is a very useful tool in model selection and in determining the theoretically possible prediction horizon for a given time series using linear modeling techniques. It should be noted that the autocorrelation decay rate for the residuals is much faster than that for the temperatures, making that signal harder to predict. For example, for a time lag of 20 min, the one used for most of our predictions, the measured temperature signal has an autocorrelation of around 0.7, whereas that for the residuals is only around 0.5. Hence, in selecting an "optimum" prediction horizon for AR data-driven models, one needs to consider the desired model accuracy, the autocorrelation function of the signal, and the inertia of the physiological process being modeled. The results of this study indicate that we can conservatively use data-driven models for making predictions up to 20–30 min ahead, which provide sufficient time for preventive actions.

The first-principles SCENARIO model, on the other hand, is driven by macroscopic energy conservation equations, and, hence, is not affected by the lack of long-term correlations in the core temperature signals. Thus, SCENARIO, as one would expect by its design, should be used for mission-planning purposes beyond 30–40 min, where the data-driven models are not capable of producing meaningful predictions. Because this first-principles model does not use past core temperature measurements as inputs, it is less susceptible to sensor failure or noisy measurements and can be used when no core temperature measurements are available.

Another important finding is that, whatever data-driven model is used for core temperature prediction, the model has to be regularized to produce credible estimates. The regularized models are especially relevant when a relatively small number of samples are available for training the model. In this case, application of parameter identification technique without regularization will lead to statistically unreliable autoregression coefficients, and as a result, to erratic predictions. The benefits of regularization should be expected when the training and testing individuals have different noise levels and, in particular, if the individual's test data are noisier than the training data. In this case, the unregularized predictions will diverge from the true temperature because noise will be amplified.

The prediction of physiological variables should also be accompanied by a measure of reliability, e.g., error bounds, about the predictions. In this respect, prediction intervals, either analytical ones or through the statistical bootstrap method, can be incorporated into data-driven and hybrid models [12]. The estimation of the reliability for first-principles model predictions is less straightforward, requiring Monte Carlo simulations.

IV. Conclusion

AR models can be developed to accurately predict core temperature in humans for up to 20–30 min ahead. The other two models tested (the first-principles SCENARIO model and a parallel hybrid model) show no advantage in terms of model fidelity over the AR model for short-term predictions. In addition, the AR model can be made "portable" from individual to individual and across studies, which offers significant advantages in real-world applications, since the same model can be "reused" for different individuals and for different environmental conditions. However, in this study, the data-driven model is only tested on a rather homogeneous population of young, healthy individuals and its portability across different demographic groups, notably different age groups, remains an open question. Also, we note that the conclusions about the superiority of the data-driven approach should be considered within the context of SCENARIO and the hybrid models used in this study. Other first-principles models may demonstrate better performance under similar conditions.

An attractive implication of the results presented in this study relate to the potential portability of data-driven models across physically fit, young athletes and soldiers performing similar types of activities. The ability to train a model on data from just a handful of individuals and use it to predict core temperature for large groups of other individuals, without the need for model tuning, will greatly facilitate the deployment of real-time physiologic monitoring and predictive systems.

Acknowledgment

The authors would like to acknowledge the support of R. Hoyt for providing the Field data, W. R. Santee for the Laboratory data, and L. G. Berglund for help with SCENARIO simulations. We also gratefully acknowledge numerous useful discussions and insights provided by T. McKenna. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

References

[1] E. E. Coris, A. M. Ramirez, and D. J. Van Durme, "Heat illness in athletes: The dangerous combination of heat, humidity and exercise," *Amer. J. Sports Med.*, vol. 34, pp. 9–16, 2004.

[2] Memorandum for Heat Injury Prevention Program. (2005, Apr. 8; 2006, Apr. 17). Department of the Army, Office of the Surgeon General, Falls Church, VA [Online]. Available: http://chppm-www.apgea.army. mil/heat/ (accessed on 11/01/2006).

[3] D. S. Moran, Y. Heled, L. Still, A. Laor, and Y. Shapiro, "Assessment of heat tolerance for post exertional heat stroke individuals," *Med. Sci. Monitor*, vol. 10, pp. CR252–257, 2004.

[4] R. F. Goldman, "Introduction to heat-related problems in military operations," in *Medical Aspects of Harsh Environments*, K. B. Pandolf and R. E. Burr, Eds., Falls Church, VA: Office of the Surgeon General, United States Army, 2001, pp. 14–21.

[5] R. W. Hoyt and K. E. Friedl, "Current status of field applications of physiological monitoring for the dismounted soldier," in *Metabolic Monitoring Technologies for Military Field Applications*, M. Poos, Ed. Washington, DC: National Academy Press, 2004, pp. 247–257.

[6] R. W. Hoyt, J. Reifman, T. S. Coster, and M. J. Buller, "Combat medical informatics: Present and future," in *Proc. AMIA Annu. Symp.*, 2002, pp. 335–339.

[7] L. G. Berglund and J. A. J. Stolwijk, "The use of simulation models of human thermoregulation in assessing acceptability of complex dynamic thermal environments," in *Energy Conservation Strategies in Buildings*, J. A. J. Stolwijk, Ed. New Haven, CT: Yale Univ, 1978, pp. 157–191.

[8] X. J. Xu and J. Werner, "A dynamic model of the human/clothing/environment-system," *J. Physiol. Anthropol. Appl. Human Sci.*, vol. 16, pp. 61–75, 1997.

[9] J. Reifman, "Alternative methods for modeling fatigue and performance," *Aviation, Space, Environ. Med.*, vol. 75, pp. A173–A180, 2004.

[10] L. Ljung, *System Identification—Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[11] M. Thompson and M. A. Kramer, "Modeling chemical processes using prior knowledge and neural networks," *Amer. Inst. Chem. Eng. J.*, vol. 40, pp. 1328–1340, 1994.

[12] N. O. Oleng', A. Gribok, and J. Reifman, "Error bounds for data-driven models of dynamical systems," *Comput. Biol. Med.*, vol. 37, pp. 670–679, 2007.

[13] N. O. Oleng' and J. Reifman, "Hybrid approaches to physiologic modeling and prediction. conference on bio-monitoring for physiological and cognitive performance during military operations," in *Proc. SPIE Symp. Defense Security*, 2005, pp. 93–203.

[14] K. K. Kraning and R. R. Gonzalez, "A mechanistic computer simulation of human work in heat that accounts for physical and physiological effects of clothing, aerobic fitness, and progressive dehydration," *J. Therm. Biol.*, vol. 22, pp. 331–42, 1997.

[15] M. Yokota, L. Berglund, W. R. Santee, M. J. Buller, and R. W. Hoyt, "Modeling physiologic responses to military scenarios: Initial core temperature and downhill work," *Aviation, Space, Environ. Med.*, vol. 76, pp. 475–480, 2005.

[16] R. A. Haslam and K. C. Parsons, "Using computer-based models for predicting human thermal responses to hot and cold environments," *Ergonomics*, vol. 37, pp. 399–416, 1994.

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

[18] D. Fiala, K. J. Lomas, and M. Stohrer, "Computer prediction of human thermoregulatory and temperature responses to a wide range of environmental conditions," *Int. J. Biometeorol.*, vol. 45, pp. 143–59, 2001.

[19] G. E. P. Box, G. M Jenkins, and G. C. Reinsel, *Time Series Analysis, Forecasting and Control*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[20] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[21] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics*. Dordrecht, The Netherlands: Reidel, 1986.

[22] A. R. Gallant, *Nonlinear Statistical Models*. New York: Wiley, 1987.

[23] D. Psichogios and L. Ungar, "A hybrid neural network-first principles approach to process modeling," *Amer. Inst. Chem. Eng. J.*, vol. 38, pp. 1499–1511, 1992.

[24] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. New York: Wiley, 1977.

[25] A. Bjorck, "Least Squares Methods," in *Handbook of Numerical Solution of Equations in Rn,* vol. 1, P. G. Ciarlet and J. L. Lions, Eds. Amsterdam, The Netherlands: North Holland/Elsevier, 1987.

[26] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, pp. 1–48, 1992.

[27] C. O'Brien, R. W. Hoyt, M. J. Buller, J. W. Castellani, and A. J. Young, "Telemetry pill measurements of core temperature in humans during active heating and cooling," *Med. Sci. Sports Exercise*, vol. 30, pp. 468–472, 1998.

[28] W. R. Santee, L. Berglund, G. A. Cardello, C. A. Winterhalter, and T. L. Endrusick "Physiological and psychological assessment of volunteers wearing air permeable battle-dress (BDU) uniforms during intermittent exercise," U. S. Army Res. Ins. Environ. Med., Natick, MA, Tech. Rep. TR06–06, 2006.

[29] V. A. Morozov, *Methods for Solving Incorrectly Posed Problems*. New York: Springer-Verlag, 1984.

[30] M. Yokota, W. Matthew, L. Berglund, M. J. Buller, and R. W. Hoyt, "Characterization of uncertainties in a thermal strain prediction model for military applications," U.S. Army Res. Inst. Environ. Med. Natick, MA, Tech. Rep. T04–03, 2003.

[31] A. Gribok, T. McKenna, and J. Reifman, "Regularization of body core temperature prediction during physical activity," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc. Eng. Revolution Biomed.*, 2006, pp. 459–463.

Andrei V. Gribok received the B.S. degree in systems science and the M.S. degree in nuclear engineering from the Moscow Institute of Physics and Engineering, Moscow, Russia, both in 1987, and the Ph.D. degree in biological science from the Moscow Institute of Biological Physics, Moscow, in 1996.

He is a Research Assistant Professor in the Department of Nuclear Engineering, University of Tennessee, Knoxville. He is currently on an inter-personnel agreement assignment with the Telemedicine and Advanced Technology Research Center, Fort Detrick, MD. His current research interests include inverse and ill-posed problems in engineering, statistical learning, and model misspecification in statistics. He is the author or coauthor of three book chapters and numerous journal and conference papers.

Mark J. Buller received the B.Sc.(Hons.) degree in applied psychology from the University of Wales College of Cardiff, Cardiff, U.K., in 1991. He is currently working toward the Sc.M. degree in computer science at Brown University, Providence, RI.

He is currently with the U.S. Army Research Institute of Environmental Medicine, Natick, MA. He has been a lead in the development of wearable physiological monitoring systems for over ten years, and is currently a member of the North Atlantic Treaty Organization (NATO) Research Technology Group "Real-Time Physiological and Psycho-Physiological Status Monitoring for Human Protection and Operational Health Applications". His current research interests include understanding human health state in harsh environments through the use of machine learning techniques.

Jaques Reifman received the B.S. degree in civil engineering from Rio de Janeiro State University, Rio de Janeiro, Brazil, in 1980, the B.B.A. degree in business administration from Rio de Janeiro Federal University, Rio de Janeiro, in 1985, and the M.S.E. and Ph.D. degrees in nuclear engineering from the University of Michigan, Ann Arbor, in 1985 and 1989, respectively.

He is a Senior Research Scientist in the Department of the Army, U.S. Army Medical Research and Materiel Command (USAMRMC), Fort Detrick, MD, where he is also the Director of the Department of Defense Biotechnology High Performance Computing Software Applications Institute for Force Health Protection and the USAMRMC Bioinformatics Cell at the Telemedicine and Advanced Technology Research Center. His current research interests include the areas of physiological signal processing, statistical pattern recognition, artificial intelligence, data mining, biomathematical modeling, bioinformatics, genomics, and proteomics.