Comparison of Mathematical Model Predictions to Experimental Data of Fatigue and Performance

Hans P. A. Van Dongen

Van Dongen HPA. ***Comparison of mathematical model predictions to experimental data of fatigue and performance.*** **Aviat Space Environ Med 2004; 75(3, Suppl.):A15–36.**

As part of the "Fatigue and Performance Modeling Workshop," six modeling teams made predictions for temporal profiles of fatigue and performance in five different scenarios. One scenario was based on a laboratory study of fatigue and performance during 88 h of extended wakefulness with or without nap opportunities. Another scenario was based on a field study of alertness in freight locomotive engineers. Two scenarios were based on laboratory studies with various conditions of chronic sleep restriction and recovery. There was a theoretical scenario for future ultra-long-range flight operations as well. Experimental data were available for all scenarios except the latter. The model predictions were compared with the experimental data; after linear scaling using mixed-effects regression, mean square errors were computed to quantify goodness-of-fit. The six models were also compared among each other on the basis of these mean square errors. The present paper provides detailed information about the results of these comparisons. The models were capable of predicting the data for some scenarios fairly well. However, predicting the data for the two scenarios involving chronic sleep restriction was more problematic. Differences among the predictions from the six models were relatively small, suggesting that these models have a broad common basis. More experimental research is needed to yield new insights for the further development of fatigue and performance models.

Keywords: biomathematical models, goodness-of-fit, neurobehavioral performance, subjective sleepiness, sleep schedules, work schedules.

AS PART OF THE PREPARATIONS for the Fatigue and Performance Modeling Workshop held in Seattle, WA, on June 13 and 14, 2002, predictions of human fatigue and performance were solicited for four different sleep/wake/work scenarios from the mathematical modeling community in this area. Six modeling teams responded and provided model predictions for at least three of the scenarios. These predictions were then compared with experimental data available from recent experiments for three of the four scenarios. A fifth scenario was presented to the six modeling teams at the Workshop proper. Model predictions for this scenario provided at the end of the first day of the Workshop were compared overnight to the data from a recent experiment as well. The results of all the comparisons of model predictions to experimental data were presented on the second day of the Workshop. The present paper represents the contents of this presentation, expanded with some additional analyses based on the discussion that followed the presentation.

The six modeling teams and their corresponding fatigue and performance models (25) are listed in **Table I**. They are given labels A through F by which they will be referred to throughout the paper. The Workshop aimed to compare and contrast the features and capabilities of these fatigue and performance models and to identify critical gaps in fatigue and performance research (34). No modeling team had any advance knowledge about the fatigue and performance data or metrics corresponding to the scenarios, and none of the six models was developed, tested, or validated using the scenarios or their data, except where noted otherwise in the text below. The modeling teams were also not informed about literature sources pertaining to any of the scenarios prior to the Fatigue and Performance Modeling Workshop.

Key questions that the Workshop aimed to address included: "Where do current models converge?" and "What is missing from the current models?" (34). For this reason, sleep/wake/work scenarios were selected so as to be challenging to the existing fatigue and performance models. The five scenarios are presented in the next section. The subsequent section describes the statistical methodology used to compare the model predictions to the experimental data. The results of the comparisons are shown in a series of figures and tables, and these results are discussed at the end of the paper.

SCENARIOS

Scenario 1: 88 h of Extended Wakefulness With and Without Naps

The first scenario for which the modeling teams were asked to provide fatigue and performance predictions was based on a laboratory experiment in which subjects maintained wakefulness for 88 h. There were two experimental conditions. Subjects in condition 1 (n = 13) were kept awake the entire 88 h. For subjects in condition 2 (n = 12), the 88 h of extended wakefulness were interrupted by 2-h nap opportunities every 12 h. All subjects were healthy male adults (age range 21–48) living in or around Philadelphia, PA (latitude/longitude = 39.9°N/75.2°W). They had no traces of drug use

From the Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, PA.

Address correspondence to: Hans P. A. Van Dongen, M.S., Ph.D., who is a Research Assistant Professor of Sleep and Chronobiology, Unit for Experimental Psychiatry, University of Pennsylvania School of Medicine, 1019 Blockley Hall, 423 Guardian Dr., Philadelphia, PA 19104-6021; vdongen@mail.med.upenn.edu.

Reprint & Copyright © by Aerospace Medical Association, Alexandria, VA.

TABLE I. MATHEMATICAL MODELS OF FATIGUE AND PERFORMANCE.

	Model Name	Lead Author(s)	References
A	CHS Chronic Fatigue Model	M. Spencer & A. Belyavin	(9,33)
B	Circadian Alertness Simulator	M. Moore-Ede	(27,28)
C	Fatigue Audit InterDyne	D. Dawson & A. Fletcher	(12,32)
D	Interactive Neurobehavioral Model	M. Jewett & R. Kronauer	(21,22)
E	Sleep, Activity, Fatigue, and Task Effectiveness Model	S. Hursh	(19,20)
F	Sleep/Wake Predictor	S. Folkard & T. Åkerstedt	(4,17)

Six fatigue and performance models were compared and contrasted using five different sleep/wake/work scenarios for which predictions of fatigue and performance were to be made. The six models and their lead authors, as well as two key references, are listed in this table ordered alphabetically by model name. Each model is given a label (A through F) by which it is referred to throughout the paper.

in their blood or urine prior to entering the laboratory. They were good sleepers, habitually sleeping between 6 and 9 h · d^{-1} (hours/day), and they were neither extreme morning-types nor extreme evening-types. The two experimental conditions considered here are the placebo conditions of an experiment described in more detail by Van Dongen et al. (38).

Table II shows the protocol description given to the modeling teams for scenario 1. Only the baseline days and the 88 h of extended wakefulness of the experiment were considered. The experiment ended with three recovery days, which were not considered part of the scenario. During all periods of scheduled wakefulness, subjects performed intensively on a 30-min computerized test battery at 2-h intervals (beginning at 08:00, 10:00, 12:00, etc.). At all other times of wakefulness, non-vigorous activities such as light reading, watching videos, or casual conversation were allowed. Subjects were instructed to try to sleep during times in bed. The laboratory was isolated from the environment and had a near-constant temperature of 21°C. Light exposure was near-constant at about 40 lux during scheduled waking periods, and less than 1 lux during scheduled sleep. No drugs and/or stimulants (including caffeine, alcohol, tobacco) were allowed inside the laboratory.

Model predictions were requested for subjective sleepiness and for neurobehavioral performance capability at all 2-h intervals of wakefulness (at 08:00, 10:00, 12:00, etc.) throughout the scenario and for each of the two conditions. For the model to data comparisons, experimental data were available from the neurobehavioral assessment battery administered at these time points. For subjective sleepiness, Karolinska Sleepiness Scale (KSS) (5) ratings were obtained near the end of each neurobehavioral test bout. For neurobehavioral performance capability, the number of lapses on a 10-min psychomotor vigilance task (PVT) (16) administered near the beginning of each test bout was used. **Fig. 1** shows the data acquired for scenario 1. The baseline period of the scenario was not used for the model to data comparisons, and is omitted from the figure.

Scenario 2: 14 d of Partial Sleep Deprivation

The second scenario for which the modeling teams were asked to provide fatigue and performance predictions was based on a laboratory experiment in which subjects were partially sleep deprived for 14 d. Two experimental conditions were considered. For subjects in condition 1 (n = 13), sleep was restricted to 4 h · d^{-1} (03:30–07:30) for 14 d. For subjects in condition 2 (n = 11), sleep was restricted to 6 h · d^{-1} (01:30–07:30) for 14 d. Subjects were healthy male and female adults (age range 21–38) living in or around Philadelphia, PA. They had no traces of drug use in their blood or urine prior to entering the laboratory. They were good sleepers, habitually sleeping between 6.5 and 9 h · d^{-1}, and they were neither extreme morning-types nor extreme evening-types. The experiment is described in more detail by Van Dongen et al. (35,37).

Table III shows the protocol description given to the modeling teams for scenario 2. During all periods of scheduled wakefulness, subjects performed intensively on a 30-min computerized test battery at 2-h intervals (beginning at 07:30, 09:30, 11:30, etc.). At all other times of wakefulness, non-vigorous activities such as light reading, watching videos, or casual conversation were allowed. Subjects were instructed to try to sleep during times in bed. The laboratory was isolated from the environment and had a near-constant temperature of 21°C. Natural daylight entered the laboratory during scheduled wakefulness. Waking light exposure was, therefore, variable depending on outside light conditions (daylight saving time was in effect during the summer months). Light exposure inside the laboratory did not exceed approximately 100 lux during scheduled waking periods, and was less than 1 lux during scheduled sleep. No drugs and/or stimulants (including caffeine, alcohol, tobacco) were allowed inside the laboratory.

Model predictions were requested for subjective sleepiness and for neurobehavioral performance capa-

TABLE II. PROTOCOL DESCRIPTION FOR SCENARIO 1.

17:00–23:30	Baseline wakefulness period #1
23:30–07:30	Scheduled time in bed for baseline sleep #1
07:30–23:30	Baseline wakefulness period #2
23:30–07:30	Scheduled time in bed for baseline sleep #2
07:30–23:30	Baseline wakefulness period #3
23:30–07:30	Scheduled time in bed for baseline sleep #3
07:30	*Condition 1:* 88 h of continuous wakefulness
	Condition 2: 88 h of wakefulness except for seven 2-h naps scheduled at 12-h intervals (14:45–16:45, 02:45–04:45)

This protocol description was given to the modeling teams for scenario 1. Subjects were randomized to one of two different conditions; model predictions were solicited for both conditions. The entire scenario was 150.5 h long. It began at 17:00, and ended 6.3 d later at 23:30 at the end of the 88 h of extended wakefulness.

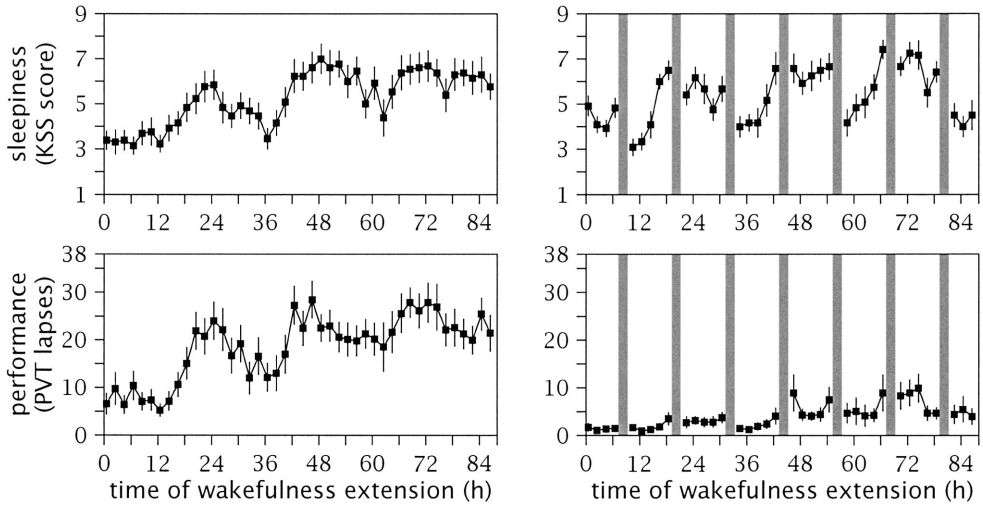

Fig. 1. Experimental data for scenario 1. The data were taken from a laboratory experiment in which subjects maintained wakefulness for 88 h. There were two experimental conditions: subjects in condition 1 (left panels) were kept awake the entire 88 h; subjects in condition 2 (right panels) received 2-h nap opportunities (gray bars) every 12 h. Subjective sleepiness data (top panels) were obtained with the Karolinska Sleepiness Scale (KSS) which yielded self-ratings ranging from 1 ("very alert") to 9 ("very sleepy"). Neurobehavioral performance data (bottom panels) were obtained with a 10-min psychomotor vigilance task (PVT) for which the number of lapses (reaction times longer than 500 ms) was counted. Group mean data are shown, with error bars indicating standard errors of the mean. Upwards corresponds to greater sleepiness or worse performance in all four panels. The abscissa shows time (in hours) since awakening from the last baseline sleep period. The data for condition 1 clearly show the build-up of sleepiness and performance impairment with progressing time awake; in addition, there is considerable circadian rhythmicity in the data. In condition 2, these temporal dynamics are dampened. The differences between the two conditions observed in subjective sleepiness and neurobehavioral performance capability during the first few hours of the scenario, before the conditions became experimentally distinct, reflect natural variability among individuals.

bility, at all 2-h intervals of wakefulness (at 07:30, 09:30, 11:30, etc.) throughout the scenario and for each of the two conditions. For the model to data comparisons, experimental data were available from the neurobehavioral assessment battery administered at these time points. For subjective sleepiness, Karolinska Sleepiness Scale (KSS) ratings were obtained near the end of each neurobehavioral test bout. For neurobehavioral performance capability, the number of lapses on a 10-min psychomotor vigilance task (PVT) administered near the beginning of each test bout was used. **Fig. 2, 3, and 4** show the data acquired for scenario 2. The baseline period of the scenario was not used for the model to data comparisons, and is omitted from the figures. As part of a data exchange agreement made several years ago, the authors of model D received PVT data for five subjects from this experiment prior to the Workshop. These data were reportedly not used to update the model prior to the Workshop.

TABLE III. PROTOCOL DESCRIPTION FOR SCENARIO 2.

17:00–23:30		Baseline wakefulness period #1
23:30–07:30		Scheduled time in bed for baseline sleep #1
07:30–23:30		Baseline wakefulness period #2
23:30–07:30		Scheduled time in bed for baseline sleep #2
07:30–23:30		Baseline wakefulness period #3
23:30–07:30		Scheduled time in bed for baseline sleep #3
Condition 1	*Condition 2*	
07:30–03:30	07:30–01:30	Extended wakefulness period #1
03:30–07:30	01:30–07:30	Restricted sleep opportunity #1
⋮	⋮	⋮
07:30–03:30	07:30–01:30	Extended wakefulness period #14
03:30–07:30	01:30–07:30	Restricted sleep opportunity #14
07:30–23:30		Pre-recovery wakefulness period
23:30–07:30		Scheduled time in bed for recovery sleep #1
07:30–23:30		Recovery wakefulness period #1
23:30–07:30		Scheduled time in bed for recovery sleep #2
07:30–23:30		Recovery wakefulness period #2
23:30–07:30		Scheduled time in bed for recovery sleep #3
07:30–10:30		Recovery wakefulness period #3

This protocol description was given to the modeling teams for scenario 2. Subjects were randomized to one of two different conditions; model predictions were solicited for both conditions. The entire scenario was 473.5 h long. It began at 17:00, and ended 19.7 d later at 10:30 after the third night of recovery sleep.

Scenario 3: Freight Locomotive Engineers on the Extra Board

The third scenario for which the modeling teams were asked to provide predictions was based on field data collected from experienced freight locomotive engineers on the extra board. Extra board engineers are on call and, therefore, are exposed to irregular and somewhat unpredictable work schedules (within the limits of U.S. service legislation). They are typically notified of their next run at least 2 h prior to the report time. When off duty, they select their wake and sleep periods at their own discretion. Times in bed are often spent away from home due to the duration of the runs.

Train driving is a non-vigorous, highly cognitive activity that generates considerable mental workload from continuous mental calculations, spatial memory usage, and vigilance monitoring. Depending on the model of locomotive and the train speed, the noise level can be as low as 75 dBA or in excess of 85 dBA. Light

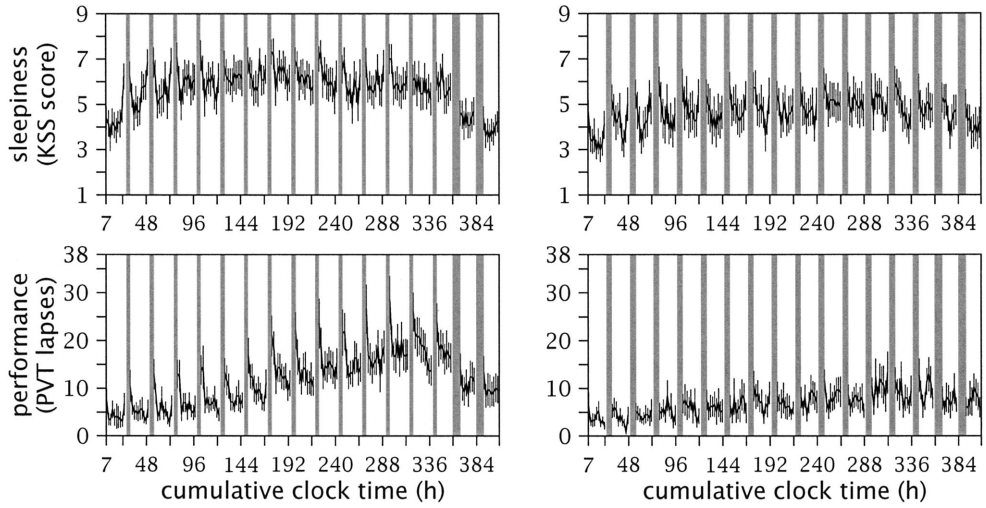

Fig. 2. Experimental data for scenario 2. The data were taken from a laboratory experiment in which subjects were partially sleep deprived for 14 d. There were two experimental conditions: subjects in condition 1 (left panels) were restricted to 4 h sleep (03:30–07:30) per day; subjects in condition 2 (right panels) were restricted to 6 h sleep (01:30–07:30) per day. Following the 14-d restriction period, subjects in both conditions received 8-h recovery sleep opportunities (23:30–07:30) for 3 d (only 2 recovery days are shown). All sleep periods are marked with gray bars. The abscissa shows cumulative clock time (in hours). Other details are the same as for Fig. 1.The data for both conditions of scenario 2 display a build-up of sleepiness and performance impairment over days of sleep restriction, and reduction thereof over the recovery days. Circadian variation is seen within days. In addition, the data collected immediately on awakening tend to show sleep inertia.

exposure can be as high as 10,000 lux during day running when heading into the sun, or as low as 100 lux during day running away from the sun. At night engineers generally dim the panel lights, creating an environment with less than 2 lux of light.

A total of 10 engineers, whose trajectory ran from Whitefish, MT (latitude/longitude = 48.4°N/114.3°W) to Havre, MT (48.6°N/109.7°W) and back, kept a paper log over 14 consecutive days in the period from February until April, 1994. Bed times and work times were gathered from these logs. The 10 sleep/wake/work schedules thus collected were all different. Together they constituted the scenario presented to the modeling teams. The engineers were healthy male adults (age range 36–54), who were randomly tested for drugs and alcohol. The modeling teams were asked to assume that the engineers were not under the influence of any such substances while on duty, although caffeine was consumed ad libitum. Napping was not allowed during scheduled work periods. No reported stressful life events occurred for any of the engineers during and just prior to the data collection period.

Model predictions were requested for subjective alertness at all 1-h intervals of wakefulness (00:00, 01:00, 02:00, etc.) for each of the 10 engineers. Daylight saving time was in effect for engineers #2, #5, #6, and #9. Ambient temperature could range from 15°C to 35°C. The engineers lived an average of 7 mi away from the home terminal and commuted 15–30 min each direction. For the model to data comparisons, experimental data were available from a 4-point alertness scale (1: "fully alert"; 2: "moderately alert"; 3: "drowsy"; 4: "fighting sleep") on which the engineers were asked to rate themselves at regular intervals during work periods. **Fig. 5** gives an overview of the scenario and the available data for the 10 engineers. Three of the engineers never reported a 4 ("fighting sleep") during the 14-d recording period; and one of them (engineer #4)

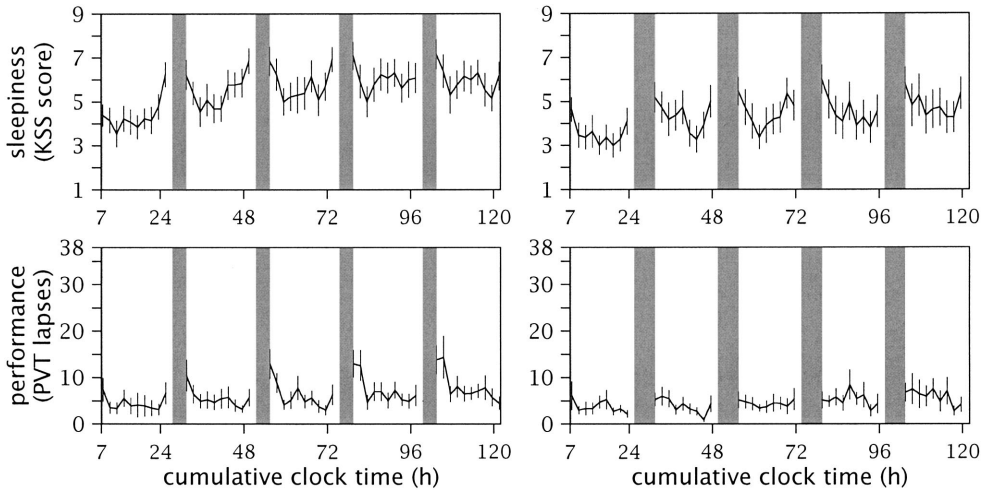

Fig. 3. Enlargement of the first 5 d of experimental data for scenario 2. The panels correspond to those in Fig. 2 but show only cumulative clock times 7 (07:00) through 122 (02:00, 4.8 d later), so as to enhance the discernibility of data points within days. Other details are the same as for Fig. 2.

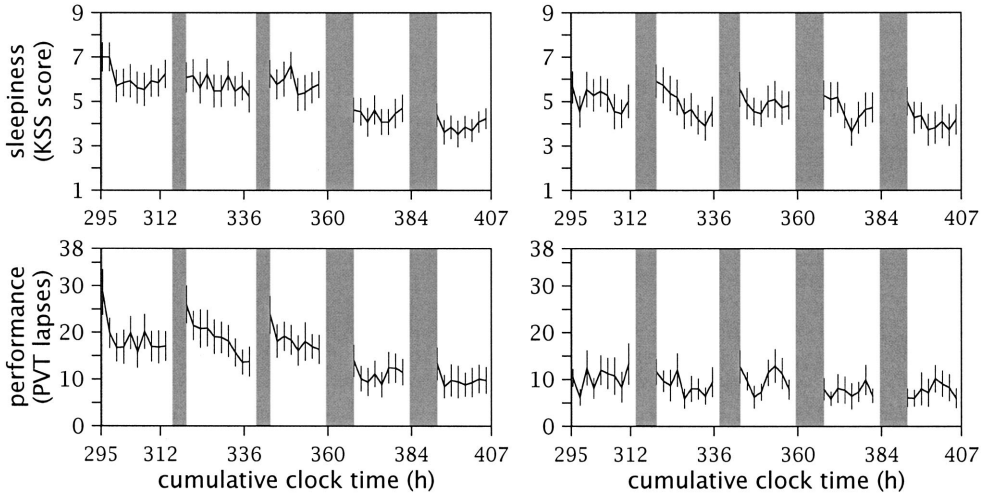

Fig. 4. Enlargement of the last 5 d of experimental data for scenario 2. The panels correspond to those in Fig. 2 but show only cumulative clock times 295 (07:00) through 407 (23:00, 4.7 d later). Other details are the same as for Fig. 2 and 3.

never reported a 3 ("drowsy") either. The data set is described in more detail by Pollard (30).

Scenario 4: Ultra-Long-Range Flight Operations

The fourth scenario for which the modeling teams were asked to provide predictions was based on a theoretical schedule for ultra-long-range (ULR) flight operations involving four crewmembers. This scenario was designed to be a realistic protocol for a possible future ULR great circle (polar route) connection between JFK airport, New York, NY (latitude/longitude = 40.6°N/74.8°W) and HKG airport, Hong Kong, China (22.3°N/113.9°E). The theoretical flight from JFK airport to HKG airport would take 20 h and the theoretical return flight would take 18 h, with a 50-h layover in Hong Kong scheduled in between. A total of 13 time zones would be crossed during each flight. **Table IV** summarizes the protocol description given to the modeling teams, involving different sleep/wake/work schedules for each of the four pilots. The theoretical pilots would be healthy, non-smoking male adults (age range 23–60) living in the urban area surrounding JFK airport. They would be good to fair sleepers with no sleep disorders, and they would have no extreme morningness or eveningness. They would be moderate habitual coffee drinkers. Other than caffeine, they would have no traces of drug use in their blood or urine prior to the flights.

Model predictions were requested for subjective sleepiness and neurobehavioral performance capability at all 1-h intervals (00:00, 01:00, 02:00, etc.) for each of the four pilots, assuming an ULR flight departure date of March 26. Specific data for hypothesized caffeine consumption were provided as part of the scenario; the modeling teams were requested to use this information for their predictions, but they could choose to ignore it. Light exposure was projected to be as follows: 600 lux during daytime (i.e., sunrise to sunset) wake periods

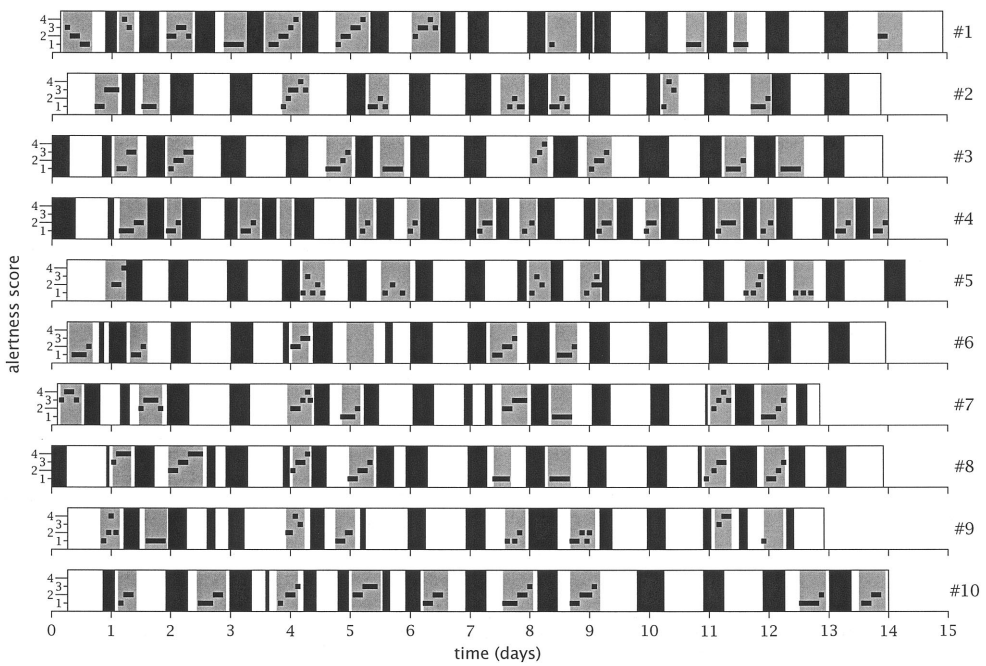

Fig. 5. Scenario overview and experimental data for scenario 3. The scenario and the data were derived from a field experiment involving 10 experienced freight locomotive engineers on the extra board. They were exposed to irregular and somewhat unpredictable work times. They kept track of their sleep/wake/work schedules for 14 d. The timeline bars in the figure display the occurrence of sleep (black), wakefulness (white), and work (gray) periods for each of the 10 engineers (labeled #1 through #10). The abscissas show cumulative time (in days). During work times, subjective alertness data were obtained with an alertness scale that yielded self-ratings ranging from 1 ("very alert") to 4 ("fighting sleep"). The alertness scores used as data for scenario 3 are marked in the figure with black points (which frequently overlap due to the condensed time scale). The vertical axis on each timeline bar indicates the alertness scale (1–4); upwards corresponds to greater sleepiness.

TABLE IV. SUMMARIZED PROTOCOL DESCRIPTION FOR SCENARIO 4.

07:00–20:00	13-h pre-departure period
07:00	Awakening from sleep at home
15:00–17:00	Pre-flight nap period pilots 2 and 4
16:00–17:00	Pre-flight nap period pilots 1 and 3
19:00	Duty starts at JFK
20:00–16:00	20-h flight from JFK to HKG
20:00	Departure from JFK
21:00–00:00	In-flight rest period pilot 3
21:00–03:45	In-flight rest period pilot 4
04:00–07:00	In-flight rest period pilot 1
04:00–10:00	In-flight rest period pilot 2
11:00–11:30	In-flight rest period pilot 3
11:00–12:00	In-flight rest period pilot 4
13:00–13:30	In-flight rest period pilot 1
13:00–14:00	In-flight rest period pilot 2
16:00	Arrival at HKG
16:00–18:00	50-h layover period
17:00	Duty complete
18:00	Arrival at layover hotel
19:00–21:00	Layover nap period pilot 1
19:00–23:00	Layover nap period pilot 2
23:00–01:00	Layover nap period pilot 3
23:00–03:00	Layover nap period pilot 4
11:00–15:00	Layover sleep period pilots 1 and 3
11:00–18:00	Layover sleep period pilots 2 and 4
09:00–14:00	Layover sleep period pilots 1 and 3
09:00–15:00	Layover sleep period pilots 2 and 4
17:00	Duty starts at HKG
18:00–12:00	18-h flight from HKG to JFK
18:00	Departure from HKG
19:00–19:30	In-flight rest period pilot 3
19:00–20:00	In-flight rest period pilot 4
21:00–21:30	In-flight rest period pilot 1
21:00–22:00	In-flight rest period pilot 2
23:00–02:00	In-flight rest period pilot 3
23:00–04:00	In-flight rest period pilot 4
05:00–08:00	In-flight rest period pilot 1
05:00–10:00	In-flight rest period pilot 2
12:00	Arrival at JFK
12:00–22:00	10-h post-departure wakefulness
13:00	End of duty
22:00–08:00	10-h recovery sleep period

This is a summary of the protocol description given to the modeling teams for scenario 4. All times are given in New York, NY local time. The entire scenario was 121.0 h long. It began at 07:00 with awakening on the day of departure from JFK airport, included a 2-d layover after arrival at HKG airport, included the return flight, and ended at 08:00 just over 5 d after it began. Sleep times were different for each of the four crewmembers. The scenario also provided specific information about caffeine consumption. Since none of the modeling teams used this information, however, it is not reproduced here.

and 5 lux during daytime sleep periods on the ground; 40 lux during nighttime wake periods and 0 lux during nighttime sleep periods on the ground; 1500 lux on the flight deck and 5 lux in the bunk during daytime in-flight periods; and 1 lux on the flight deck and 5 lux in the bunk during nighttime in-flight periods. The modeling teams were encouraged to provide more precise predictions of (natural and artificial) light exposure as part of the modeling process. Daylight saving time would be in effect at all locations. Ambient temperature was projected to be 4°C in the JFK airport area, and 23°C in the HKG airport area as well as in the aircraft.

The fatigue and performance modeling teams were asked to assume that during the 7 d prior to the flight scenario, pilots 1 and 2 would sleep 8 h (23:00–07:00) each night and pilots 3 and 4 would sleep 6 h (01:00–07:00) each night. In addition, all crewmembers would be entrained to the local day at JFK airport for at least 2 d prior to the flight scenario. Since the ULR connection between New York and Hong Kong does not currently exist, there were no experimental data to compare the model predictions with. Rather, this scenario aimed to expose differences among the fatigue and performance models without reference to existing data.

Scenario 5: 7 d of Sleep Restriction Followed by 3 d of Recovery

The final scenario for which the modeling teams were asked to provide predictions was based on a laboratory experiment in which subjects were partially sleep deprived for 7 d. The experiment resembled the one on which scenario 2 was based, but differed in a number of details. Two experimental conditions were considered in scenario 5. For subjects in condition 1 (n = 16), sleep was restricted to 7 h · d^{-1} (00:00–07:00) for 7 d. For subjects in condition 2 (n = 18), sleep was restricted to 3 h · d^{-1} (04:00–07:00) for 7 d. In both conditions, the 7-d restriction period was followed by 3 recovery days with sleep scheduled from 23:00 until 07:00. In condition 2, neurobehavioral recovery was incomplete after the 3 recovery days, which made the recovery phase of this experiment particularly interesting. Subjects were healthy male and female adults (age range 21–62) living in or around Washington, DC (latitude/longitude = 39.4°N/76.6°W). They had no traces of drug use in their blood or urine prior to entering the laboratory. They were good sleepers with no subjective sleep or sleepiness complaints, habitually sleeping between 6 and 9 h · d^{-1}. The experiment is described in more detail by Balkin et al. (6) and Belenky et al. (8).

Table V shows the protocol description given to the modeling teams for this scenario. A cognitive performance test battery and a multiple sleep latency test were administered to the subjects four times a day. Free time for personal hygiene, meals, watching videos, or casual conversation was scheduled at 07:05–07:30, 08:30–09:00, 12:40–13:30, 14:25–15:00, 17:15–19:30, and 20:25–21:00 each day. Subjects were instructed to try to sleep during scheduled times in bed. Light exposure was variable depending on outside light conditions (daylight saving time was in effect during the summer months), but bedrooms were darkened during scheduled times in bed. The ambient temperature in the laboratory was near-constant at about 21°C. No drugs and/or stimulants (incl. caffeine, alcohol, tobacco) were allowed inside the laboratory and during the week prior to the beginning of the scenario.

Scenario 5 was first presented to the six modeling teams at the Fatigue and Performance Modeling Workshop proper. This was done to get a sense of how the models would perform during real-time end-user deployment. Model predictions were requested by the end of the first day of the Workshop. Predictions were solicited for neurobehavioral performance capability at 09:30, 12:30, 15:30, and 21:30 for all days of the scenario and for each of the two conditions. For the model to data comparisons, data were available from a 10-min psychomotor vigilance task (PVT) (16) administered at

TABLE V. PROTOCOL DESCRIPTION FOR SCENARIO 5.

10:00–23:00		Baseline wakefulness period #1
23:00–07:00		Scheduled time in bed for baseline sleep #1
07:00–23:00		Baseline wakefulness period #2
23:00–07:00		Scheduled time in bed for baseline sleep #2
07:00–23:00		Baseline wakefulness period #3
23:00–07:00		Scheduled time in bed for baseline sleep #3
Condition 1	*Condition 2*	
07:00–00:00	7:00–04:00	Extended wakefulness period #1
00:00–07:00	4:00–07:00	Restricted sleep opportunity #1
⋮	⋮	⋮
07:00–00:00	7:00–04:00	Extended wakefulness period #7
00:00–07:00	4:00–07:00	Restricted sleep opportunity #7
07:00–23:00		Pre-recovery wakefulness period
23:00–07:00		Scheduled time in bed for recovery sleep #1
07:00–23:00		Recovery wakefulness period #1
23:00–07:00		Scheduled time in bed for recovery sleep #2
07:00–23:00		Recovery wakefulness period #2
23:00–07:00		Scheduled time in bed for recovery sleep #3
07:00–21:40		Recovery wakefulness period #3

This protocol description for scenario 5 was given to the modeling teams at the Workshop. Subjects were randomized to one of two different conditions; model predictions were solicited for both conditions. The entire scenario was 323.7 h long. It began at 10:00, and ended 13.5 d later at 21:40 after the third night of recovery sleep.

these time points and analyzed to assess mean reaction time (in milliseconds). **Fig. 6** shows the data acquired for scenario 5. The baseline period of the scenario was not used for the model to data comparisons, and is omitted from the figure. The modeling teams were not informed about the use of PVT data for model to data comparisons until the second day of the Workshop. However, the data from the experiment on which the scenario was based were available to modeling team E prior to the Workshop. Their model was optimized using those data; therefore, it was not included in the model to data comparisons for this scenario.

METHODS

The fatigue and performance modeling teams were not a priori informed about the nature of the experimental data to which their model predictions would be compared. As a consequence, the model predictions for each of the five scenarios were typically provided in a different metric than the experimental data. Thus, in order to compare the models to the data, the predictions had to be scaled to project them onto the same metric as the data. **Fig. 7** illustrates this problem.

There are many ways to solve the scaling problem, each with particular advantages and disadvantages. No matter what method is used, however, scaling affects the results of the subsequent model to data comparisons. The impact of the selected scaling method must be commensurate with the intention of the comparisons. The intention was to assess goodness-of-fit, in the least-squares statistical sense, of the model predictions to the data. For this reason, a least-squares-optimal scaling approach was taken. This ensured that each of the six fatigue and performance models was scaled optimally with respect to the subsequent comparison to the data. The outcome of the comparison was, therefore, not negatively affected by the scaling procedure or differentially advantageous for specific models.

Given the availability of experimental data from multiple subjects for scenarios 1, 2, 3, and 5 (and no data at all for scenario 4), a statistically optimal scaling method taking into account multiple subjects at once was preferred. Scaling of the model predictions to the data of each individual subject without regard to the rest of the data available for the scenario would be undesirable, since the degrees of freedom involved in the scaling would have unreasonable proportions. However, the sampling times for the data in scenario 3 were different for each individual subject. Thus, scaling methods requiring multiple data values per time point to determine the scaling parameters could not be used. Further, large inter-individual differences in factors influencing fatigue and performance have been documented (35,39). The scaling method of choice, therefore, should

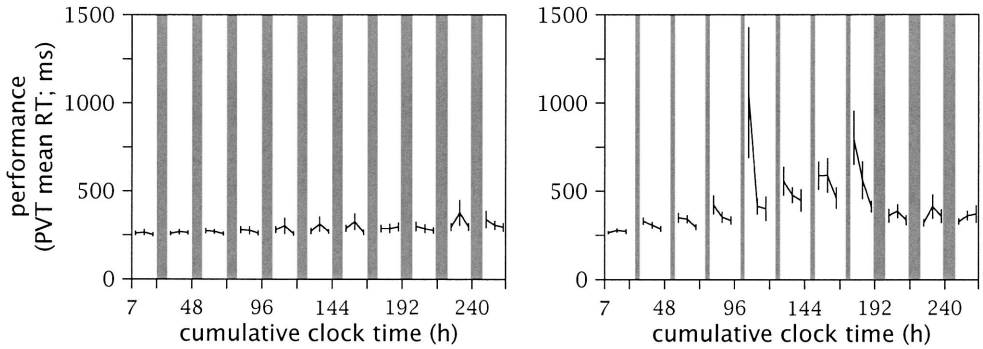

Fig. 6. Experimental data for scenario 5. The data were taken from a laboratory experiment in which subjects were partially sleep deprived for 7 d, after which they had 3 d of recovery. There were two experimental conditions: subjects in condition 1 (left-hand panel) were restricted to 7 h sleep (00:00–07:00) per day; subjects in condition 2 (right-hand panel) were restricted to 3 h sleep (04:00–07:00) per day. Following the 7-d restriction period, subjects in both conditions received 8-h sleep opportunities (23:00–07:00) for 3 d. During wakefulness, neurobehavioral performance data were obtained with a 10-min psychomotor vigilance task (PVT) for which the mean reaction time (RT; in ms) was computed. Group mean data are shown, with error bars indicating standard errors of the mean. Upwards corresponds to worse performance in both panels. The abscissa shows cumulative clock time (in hours) since awakening from the last baseline sleep period. Sleep periods are marked with gray bars. The data for condition 2 display a substantial build-up of performance impairment over the 7 d of sleep restriction, followed by a gradual but incomplete return to baseline values over the 3 recovery days. In condition 1, such temporal dynamics are nearly absent.

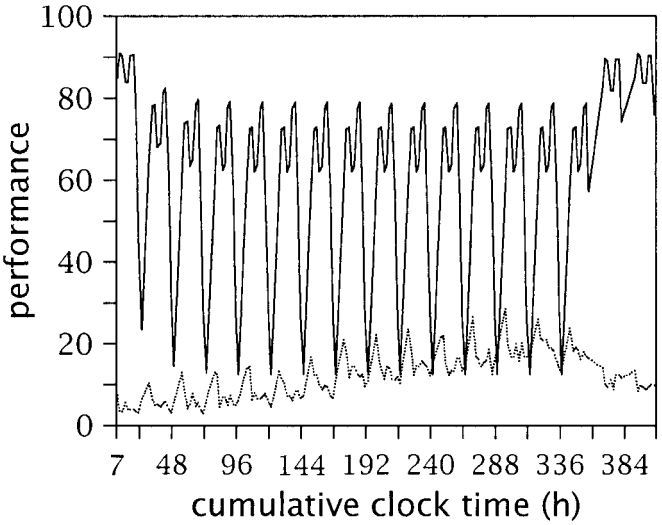

Fig. 7. Illustration of the need for scaling. The modeling teams did not know in advance the precise metrics of the experimental data to which their predictions would be compared. Consequently, model predictions typically employed different metrics than the experimental data. This is illustrated in the figure, which shows experimental data (PVT lapses; dotted curve) for condition 1 of scenario 2 as well as corresponding predictions from one of the models (prior to any scaling; solid curve), projected onto the same numerical scale (ordinate). Clearly, the vertical scales of the experimental data and the model predictions do not match. In order to make meaningful comparisons between the model and the data, therefore, scaling of the data and the predictions to a common metric is necessary. In this case, the predictions can be scaled to the same metric as the data by a linear numerical transformation involving an inversion, a reduction in range, and a vertical shift. The abscissa in the figure shows cumulative clock time (in hours).

properly distinguish between-subjects variance from within-subjects variance in the data.

Mixed-effects regression methods (29,36,40) meet all these conditions. An additional advantage of mixed-effects regression is the treatment of random effects as parameterized stochastic (Gaussian) distributions, making the method relatively robust to outliers in the experimental data. In order to restrict the degrees of freedom involved in the scaling, linear mixed-effects regression was selected as the most preferable scaling approach. (With non-linear mixed-effects regression the scaling tends to become too flexible, and the results of the model to data comparisons become difficult to interpret.) The linear mixed-effects regression model was formulated as follows:

$$y_{it} \sim \alpha + \beta x_t \qquad \text{Eq.1}$$

where $\sim$ stands for "is modeled as", y_{it} denotes the empirical values at time points t for subjects i, and x_t denotes the mathematical model predictions at time points t. The linear scaling factor β and scaling offset α were estimated using all the available data y_{it} for a given scenario at once. A normally distributed random effect for α was included in the regression model. Thus, both the population mean and variance were estimated for the scaling offset. In addition, subject-specific empirical Bayes estimates α_i of the random effect were derived. For convenience of implementation, the computations were performed using PROC NLMIXED in SAS release 8.1 (SAS Institute Inc., Cary, NC).

Once the value for β and the population mean for α were estimated for a specific model and a given scenario, the model predictions x_t could be transformed into scaled model predictions x'_t projecting on the same metric as the experimental data:

$$x'_t = \alpha + \beta x_t \qquad \text{Eq.2}$$

This allowed overlaying of the model predictions and the population average of the experimental data in a single figure for graphical comparison (except in scenario 3, where no population average time series exists). Using the subject-specific values α_i, the model predictions x_t could also be scaled to subject-specific model predictions x''_{it} projecting on the same metric as the experimental data:

$$x''_{it} = \alpha_i + \beta x_t \qquad \text{Eq.3}$$

As a measure of goodness-of-fit to compare the model predictions to the experimental data, the mean square error (MSE) was then computed as:

$$\text{MSE} = \sum_i \sum_t (y_{it} - x''_{it})^2/m \qquad \text{Eq.4}$$

where m is the total number of data points available for the scenario. This particular way of summing the squared errors ensured that the subjects in scenario 3 contributed to the overall result in proportion to the number of data points they had. In the other scenarios, all subjects had the same number of data points. Note that in scenarios 1 and 2, the first condition (with the greatest amount of wakefulness) contained more data points than the second condition, thus putting somewhat greater weight on the first condition in the computation of the MSE values.

Although the MSE values could be converted to measures of "explained variance" (as discussed in Appendix A), they were instead used directly as a quantitative means of comparing the model predictions to the data for scenarios 1, 2, 3, and 5. Lower MSE values correspond to better fit with the data. Given that the absolute magnitude of MSE values depends on the metric of the experimental data, a relative measure of goodness-of-fit was derived for which the values would have a more straightforward interpretation. The idea was to express the MSE values on a percentage scale, where 0% and 100% would correspond to the best and worst possible fits, respectively. The MSE for the best possible fit (BMSE) was approximated by taking the average experimental data over all subjects y'_t as a surrogate mathematical model $\hat{x}_t$:

$$\hat{x}_t = y'_t = \sum_i y_{it}/n \qquad \text{Eq.5}$$

where n is the number of subjects; and performing the linear mixed-effects regression-based procedure described above. (Note that $\hat{x}_t$ could not be computed for scenario 3, since the population average time series does not exist for this scenario.) The MSE for the worst possible fit (WMSE) was determined by taking a horizontal line as a surrogate model and performing the linear mixed-effects regression-based procedure again. The relative root mean square error (RRMSE) was then expressed as:

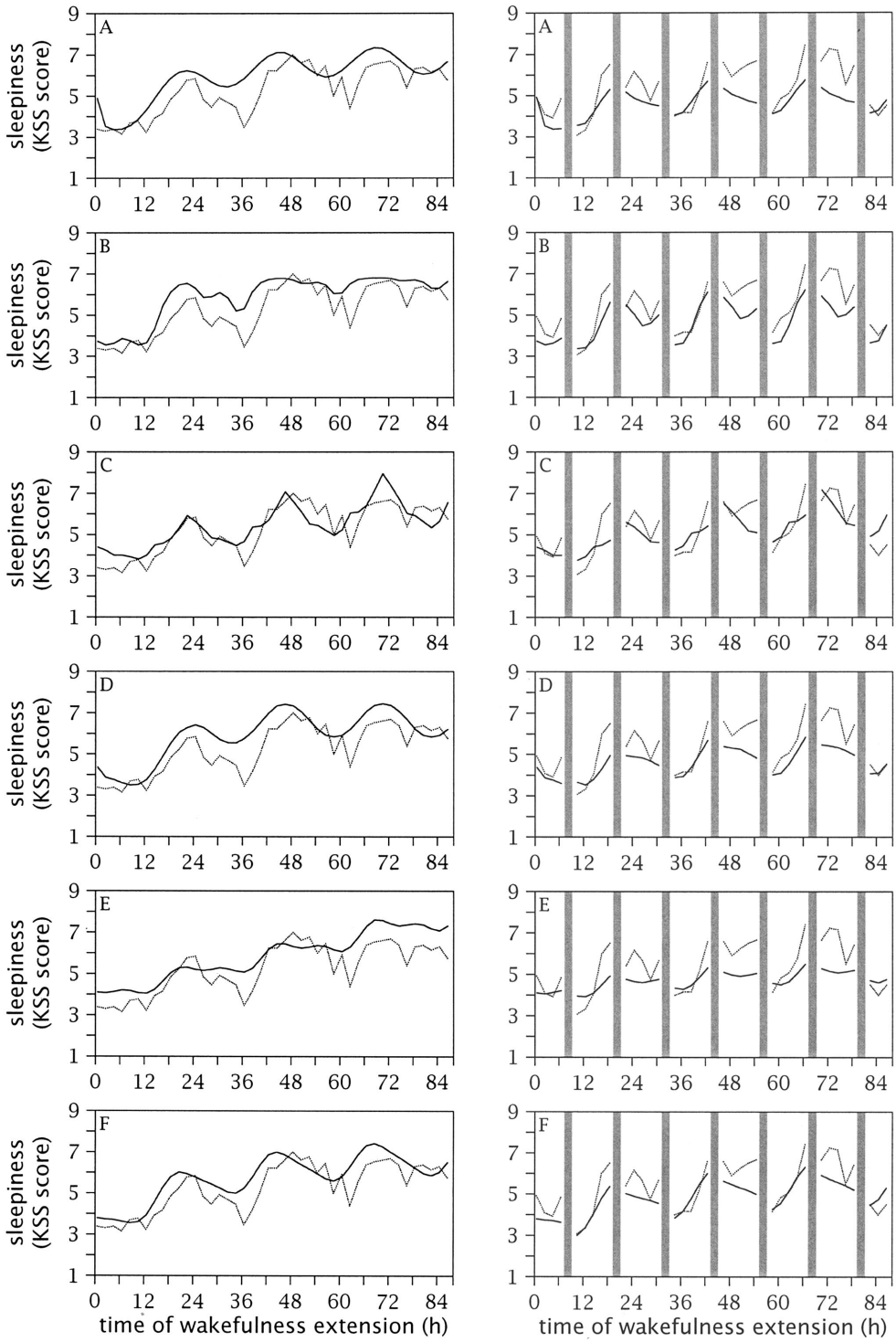

Fig. 8. Graphical comparison of models to subjective sleepiness data for scenario 1. The left-hand panels A through F show the scaled predictions (solid curves) from the corresponding models (Table I), overlaid on the average experimental data (dotted curve), for condition 1. The right-hand panels A through F show the same for condition 2. The model predictions were scaled to the data of both conditions simultaneously. The abscissa shows time (in hours) since awakening from the last baseline sleep period. The gray bars in the right-hand panels indicate 2-h nap opportunities. Refer to the main text and to Fig. 1 (top panels) for further details.

$$\mathrm{RRMSE} = 100\% \cdot \frac{\sqrt{\overline{\mathrm{MSE}}} - \sqrt{\overline{\mathrm{BMSE}}}}{\sqrt{\overline{\mathrm{WMSE}}} - \sqrt{\overline{\mathrm{BMSE}}}} \qquad \text{Eq.6}$$

By rank ordering the model predictions on the basis of the RRMSE values, models could be compared among each other (lower RRMSE values correspond to better fit with the data). The RRMSE values did not, however, allow comparison of models across scenarios, because the contextual variance is likely to vary among the scenarios.

RESULTS

Model to Data Comparisons for Scenario 1: 88 h of Extended Wakefulness

Scenario 1 was based on a laboratory experiment in which subjects maintained wakefulness for 88 h; there were two experimental conditions. **Fig. 8 and 9** show the scaled model predictions x_t' overlaid on the average data y_t' for subjective sleepiness and neurobehavioral performance capability, respectively, for each of the six

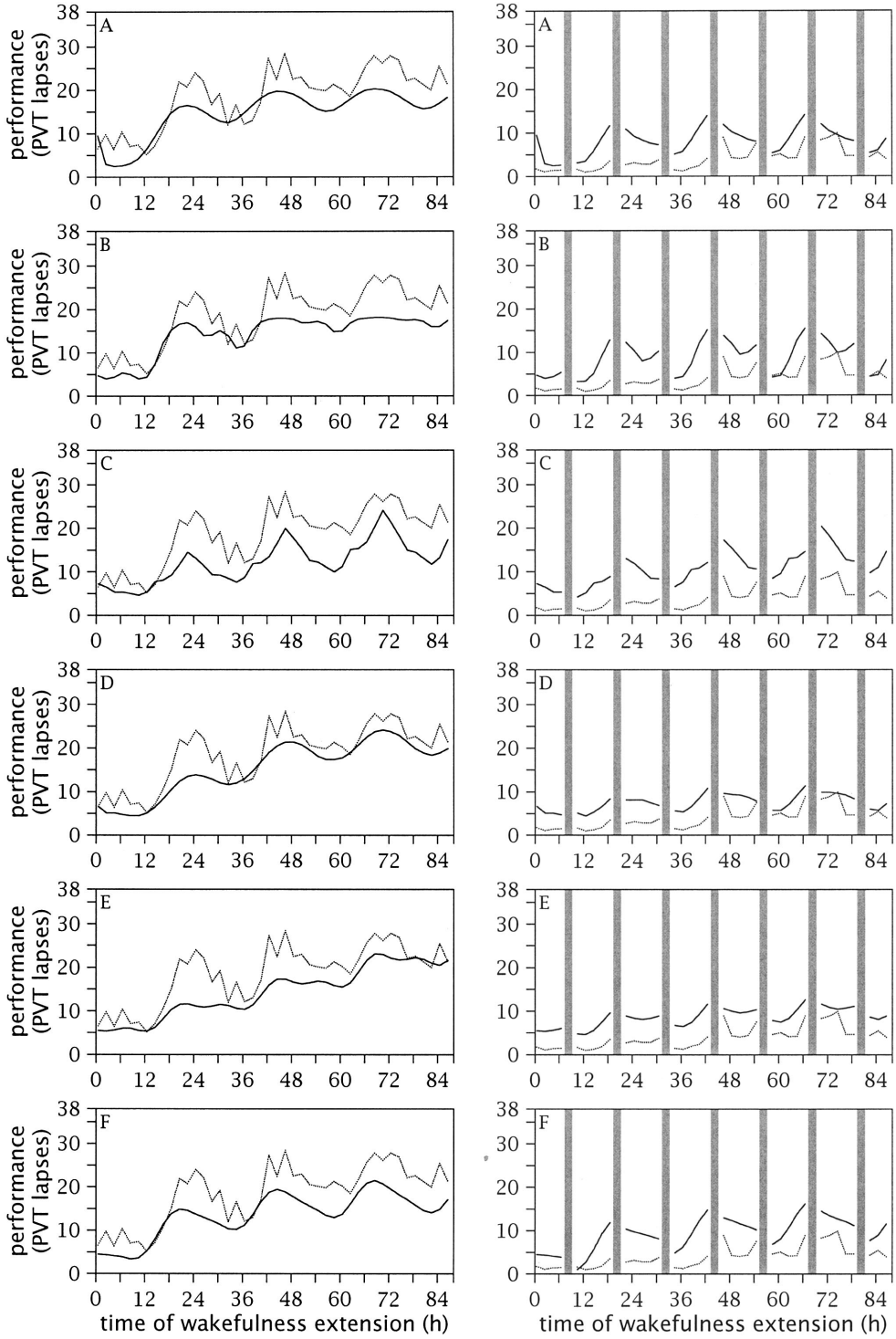

Fig. 9. Graphical comparison of models to neurobehavioral performance data for scenario 1. The left panels A through F show the scaled predictions (solid curves) from the corresponding models (Table I), overlaid on the average experimental data (dotted curve), for condition 1. The right-hand panels A through F show the same for condition 2. Other details are the same as for Fig. 8; see also Fig. 1 (bottom panels).

fatigue and performance models applied to this scenario. Linear scaling of the model predictions was performed for the two conditions in the scenario combined, so as to investigate the predictive potential of each of the models for both conditions simultaneously. The variance in the data of the combined conditions is greater than the variance in either condition alone; considering the conditions combined, therefore, yields greater potential to identify facets of the models that could be refined.

Different predictions were provided for subjective sleepiness vs. neurobehavioral performance capability for models A, D and F only. For model F these different predictions were linearly dependent, which means that the difference was trivial under linear scaling. For models B, C, and E, the same predictions were used for the comparisons to both the subjective sleepiness and the neurobehavioral performance capability data. Prior to the Fatigue and Performance Modeling Workshop, the authors of model E stated that their model was not designed to predict subjective sleepiness, and that its use for predicting subjective sleepiness should be re-

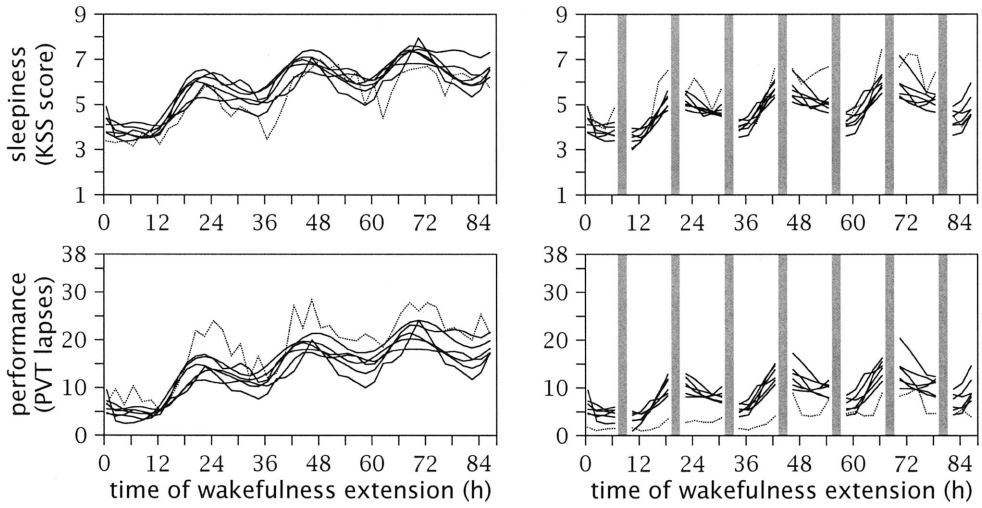

Fig. 10. Composite graphs for comparison of models to data in scenario 1. The top panels show the scaled model predictions for all six models A through F (solid curves) overlaid on the average experimental data (dotted curve) for subjective sleepiness in conditions 1 (left) and 2 (right). The bottom panels show the same for neurobehavioral performance capability in conditions 1 (left) and 2 (right). Thus, these four panels combine the graphs in Fig. 8 and 9 by collapsing the panels in these figures from top to bottom.

garded as an extension beyond the model's original scope.

Fig. 10 shows composite graphs with all six models overlaid on the average experimental data at once. This figure allows further graphical comparisons of the models to the data and to each other. It would appear that all the models systematically overestimated the subjective sleepiness data in condition 1, and underestimated these data in condition 2. The opposite appeared to be true for the neurobehavioral performance data. These appearances should be interpreted with care, however, because they are dependent on the characteristics of the selected scaling method, which was statistically optimal but not necessarily visually most favorable. In general, all graphical assessments that would change after rescaling the model predictions (i.e., stretching, shrinking, or shifting the curves vertically in the graph) should be considered conditional to the scaling method at hand. Nevertheless, the consistency by which all models underestimated the data in one condition and overestimated the data in the other condition suggests that simultaneously predicting the temporal profiles for both conditions caused the models problems.

Table VI shows the MSE and RRMSE values resulting from the quantitative comparisons of the six models to the experimental data for scenario 1, again combining the data from the two conditions. Note that the differences in the data between the two conditions (which involved different sets of subjects) during the first few hours of the scenario, before these conditions were experimentally distinct, reflected natural variability among the subjects (Fig. 1). The linear mixed-effects regression approach used to scale the model predictions to the data took between-subjects variability into account. Therefore, the MSE and RRMSE values were not unduly affected by these initial differences, which could not have been predicted by the fatigue and performance models using the information provided for the scenario.

Model to Data Comparisons for Scenario 2: 14 d of Partial Sleep Deprivation

Scenario 2 was based on a laboratory experiment in which subjects were partially sleep deprived for 14 d; there were two experimental conditions. **Fig. 11 and 12** show the scaled model predictions x_t' overlaid on the average data y_t' for subjective sleepiness and neurobehavioral performance capability, respectively, for each of the six models applied to this scenario. Linear scaling of the model predictions was performed for the two conditions in the scenario combined, so as to investigate the predictive potential of each of the models for the two conditions simultaneously. As for scenario 1, different predictions were provided for subjective sleepiness vs. neurobehavioral performance capability for models A, D, and F only. For model F these different predictions were again linearly dependent. For models B, C, and E, the same predictions were used for the comparisons to both the subjective sleepiness and the neurobehavioral performance capability data.

Fig. 13 shows composite graphs with all six models overlaid on the average experimental data at once. **Table VII** shows the MSE and RRMSE values resulting from the quantitative comparisons of the six models to the data for scenario 2, combining the data from the two conditions. Clearly, none of the models predicted the continuing build-up of subjective sleepiness and, in

TABLE VI. QUANTITATIVE COMPARISONS OF MODEL PREDICTIONS TO EXPERIMENTAL DATA FOR SCENARIO 1.

	Sleepiness		Performance	
Model	MSE	RRMSE	MSE	RRMSE
A	2.69	41.32%	63.06	36.02%
B	2.52	28.52%	63.59	38.04%
C	2.70	41.99%	62.89	35.38%
D	2.58	33.06%	59.45	22.15%
E	2.68	41.09%	61.37	29.57%
F	2.57	32.53%	61.59	30.42%

This table shows the MSE (mean square error) and RRMSE (relative root mean square error) values resulting from the statistical comparisons of the six models to the experimental data for the two conditions (combined) of scenario 1. For subjective sleepiness the data were Karolinska Sleepiness Scale (KSS) scores; for neurobehavioral performance capability the data were lapses on a psychomotor vigilance task (PVT). The results are listed by model label (see Table I) in alphabetical order. For comparison, the BMSE (estimated best possible MSE) was 2.16 and the WMSE (worst possible MSE) was 3.53 for subjective sleepiness. The BMSE was 53.90 and the WMSE was 81.11 for neurobehavioral performance capability.

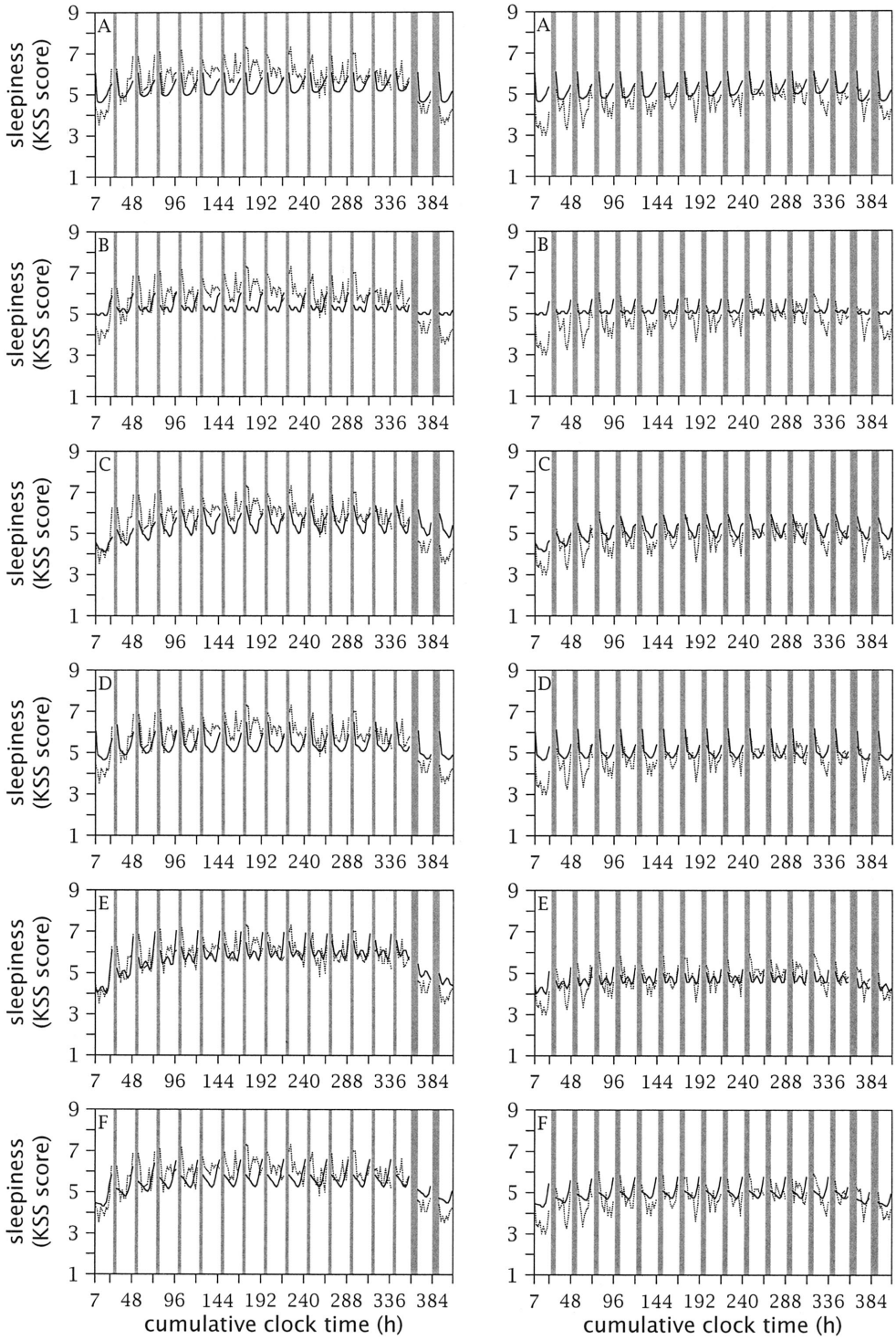

Fig. 11. Graphical comparison of models to subjective sleepiness data for scenario 2. The left panels A through F show the scaled predictions (solid curves) from the corresponding models (Table I), overlaid on the average experimental data (dotted curve), for condition 1. The right panels A through F show the same for condition 2. The model predictions were scaled to the data of both conditions simultaneously. The abscissa shows cumulative clock time (in hours); the gray bars indicate sleep periods. Refer to the main text and to Fig. 2 (top panels) for further details.

particular, performance impairment across the 14 d of sleep restriction.

It is noteworthy that the scaled predictions from model B for neurobehavioral performance capability (Fig. 12, panels B1 and B2) showed virtually no changes over time. This was not a feature of the original predictions, but a result of the statistically optimal scaling method. There was little consistency between the temporal profiles of the original predictions and the experimental data (for neurobehavioral performance capability, the linear correlation across all data points was 0.023). As a consequence, the scaling factor β corresponding to statistically optimal scaling was found to be nearly zero, which resulted in considerable blunting of temporal changes in the scaled predictions. The corresponding MSE value for goodness-of-fit was nearly identical to the WMSE value for a horizontal line (Table VII). Had the scaling factor been greater, however, the MSE value would have exceeded the WMSE value. The mathematics underlying the scaling procedure did not allow this to occur.

Just prior to the Fatigue and Performance Modeling

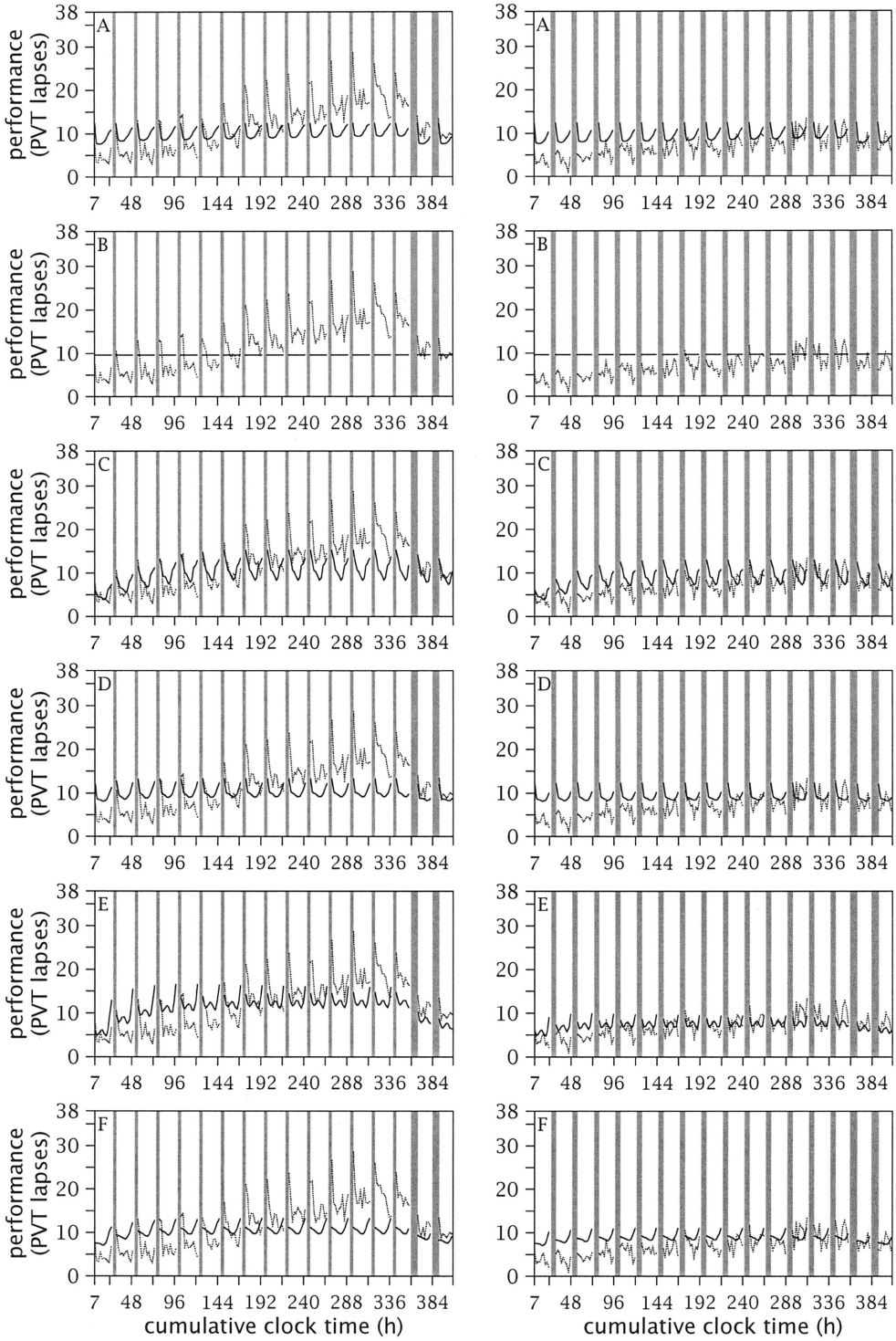

Fig. 12. Graphical comparison of models to neurobehavioral performance data for scenario 2. The left panels A through F show the scaled predictions (solid curves) from the corresponding models (Table I), overlaid on the average experimental data (dotted curve), for condition 1. The right panels A through F show the same for condition 2. Other details are the same as for Fig. 11; see Fig. 2 (bottom panels).

Workshop, model E was revised and optimized using data from scenario 5, which is similar to scenario 2. The updated predictions for neurobehavioral performance capability from this model for scenario 2 are shown in **Fig. 14**. The corresponding MSE value is 58.43, and the RRMSE value is 70.91%. This constitutes a substantial improvement with respect to the earlier predictions from this and all the other models (cf. Table VII). Still, it remained a challenge to predict the temporal profiles of both conditions of scenario 2 simultaneously. In addition, the predictions for the first few time points of the scenario differed between the two conditions, before these conditions were experimentally distinct. This discrepancy might point to a problem with the initial values used to run the model.

All six models appeared to have difficulty predicting the temporal variations within days in scenario 2. This observation can be made regardless of the influence of scaling, for it pertains to the temporal profile and not to the magnitude of changes in the predictions compared to the data. At the Workshop, it was suggested that the difficulty to properly predict variations within days might

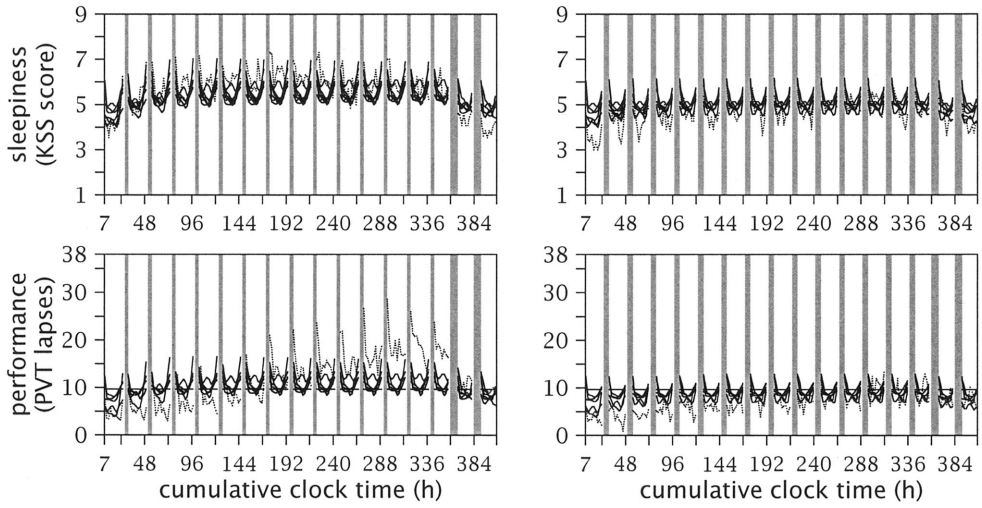

Fig. 13. Composite graphs for comparison of models to data in scenario 2. The top panels show the scaled model predictions for all six models A through F (solid curves) overlaid on the average experimental data (dotted curve) for subjective sleepiness in conditions 1 (left) and 2 (right). The bottom panels show the same for neurobehavioral performance capability in conditions 1 (left) and 2 (right). Thus, these four panels combine the graphs in Fig. 11 and 12 by collapsing the panels in these figures from top to bottom.

have to do with sleep inertia, which is the performance impairment commonly experienced right after awakening (15). It has been reported that sleep inertia largely dissipates within about 2 h after waking up (3,23). Therefore, sleep inertia could be excluded from the data by removing each data point measured immediately after a sleep period (the data were collected at 2-h intervals). It would be hypothesized that the model predictions compare more favorably to the data after removal of the data points likely to be influenced by sleep inertia. To test this hypothesis, the assessments of goodness-of-fit were repeated for neurobehavioral performance capability while specifically excluding each data point potentially affected by sleep inertia.

Table VIII shows the MSE and RRMSE values resulting from quantitative comparison of the six models, which included the revised model E, to the data of both scenario conditions following removal of all data points measured immediately after sleep. Even though Table VII (which contains the original MSE and RRMSE values) and Table VIII cannot be compared directly due to the difference in the number of data points used to assess goodness-of-fit, the results are clearly contrary to expectation. For five of the six models, goodness-of-fit was markedly reduced when data points potentially affected by sleep inertia were not included in the data set. This suggests that either sleep inertia lasted significantly longer than 2 h in the partial sleep deprivation experiment, or the overall circadian variation of sleepiness and performance in the experiment was fundamentally different than predicted by the fatigue and performance models.

Model to Data Comparisons for Scenario 3: Engineers on the Extra Board

Scenario 3 was based on field data collected from freight locomotive engineers. Each of the 10 engineers for whom data were available was on a different sleep/wake/work schedule. Engineers self-rated their alertness during their work periods, at hourly intervals or often less frequently. The model to data comparisons included only model predictions at those time points for which experimental data were available. Linear scaling of the model predictions was performed for all 10 engineers at once. Predictions were available for only five models; modeling team A did not provide any predictions for this scenario.

Since each engineer was on a different sleep/wake/work schedule, no population average time series exists for this scenario. Showing the predictions from each of the models for all 10 engineers individually, however, is not practicable. For this reason, the results for the comparison of the model predictions to the data for scenario 3 are shown only in numerical form: **Table IX** shows the MSE values resulting from the quantitative comparisons of the five models to the experimental data. RRMSE values could not be assessed, because the BMSE (as defined earlier in this paper) could not be computed due to the lack of a population average time series.

Model Predictions for Scenario 4: Ultra-Long-Range Flight Operations

Scenario 4 involved a theoretical schedule for ultra-long-range (ULR) flight operations with four crewmem-

TABLE VII. QUANTITATIVE COMPARISONS OF MODEL PREDICTIONS TO EXPERIMENTAL DATA FOR SCENARIO 2.

	Sleepiness		Performance	
Model	MSE	RRMSE	MSE	RRMSE
A	2.66	74.22%	62.60	90.76%
B	2.76	89.55%	64.60	99.99%
C	2.61	65.95%	59.19	74.57%
D	2.63	69.10%	62.92	92.27%
E	2.55	54.92%	60.57	81.21%
F	2.63	68.77%	63.27	93.87%

This table shows the MSE (mean square error) and RRMSE (relative root mean square error) values resulting from the statistical comparisons of the six models to the experimental data for the two conditions (combined) of scenario 2. For subjective sleepiness the data were Karolinska Sleepiness Scale (KSS) scores; for neurobehavioral performance capability the data were lapses on a psychomotor vigilance task (PVT). The results are listed by model label (see Table I) in alphabetical order. For comparison, the BMSE (estimated best possible MSE) was 2.23 and the WMSE (worst possible MSE) was 2.82 for subjective sleepiness. The BMSE was 44.70 and the WMSE was 64.60 for neurobehavioral performance capability.

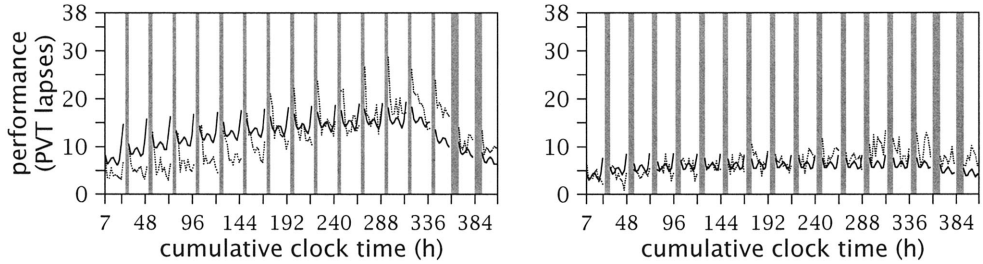

Fig. 14. Graphical comparison of revised model E to the neurobehavioral performance data for scenario 2. The left panel shows the updated, scaled model predictions (solid curve) overlaid on the average experimental data (dotted curve) for condition 1 (cf. Fig. 12, left panel E). The right panel shows the same for condition 2 (cf. Fig. 12, right panel E). Refer to the main text and to Fig. 12 for further details.

bers. No experimental data were available for this scenario. Therefore, only graphical comparisons of the models among each other could be made. (Note that statistical comparisons would require a source of variance for each time point, which the models did not yield.) No predictions were provided for this scenario by modeling teams C and D. Thus, predictions were available for only four models. None of these models took the scenario information about caffeine consumption into account.

Fig. 15 shows the predictions from models B, E, and F for each of the four crewmembers. Model predictions were requested for subjective sleepiness and neurobehavioral performance capability, but only model F distinguished between these two modalities. Furthermore, the subjective sleepiness and neurobehavioral performance predictions from model F were linearly dependent, so that the distinction was lost after linear scaling. Therefore, only subjective sleepiness predictions are shown in Fig. 15. The scaling method applied to the model predictions for the other scenarios could not be performed for scenario 4, because there were no experimental data. Instead, for each model the predictions were linearly projected onto a scale from 0 to 1, with 0 corresponding to the lowest sleepiness value and 1 corresponding to the highest sleepiness value among the combined predictions for the four crewmembers.

TABLE VIII. QUANTITATIVE COMPARISONS OF MODEL PREDICTIONS FOR SCENARIO 2 TO NEUROBEHAVIORAL PERFORMANCE DATA EXCLUDING SLEEP INERTIA.

Model	MSE	RRMSE
A	56.87	96.13%
B	57.49	99.90%
C	54.63	82.62%
D	57.42	99.46%
E	53.41	75.31%
F	56.74	95.38%

Like the two right-hand columns in Table VII, this table shows the MSE (mean square error) and RRMSE (relative root mean square error) values resulting from statistical comparisons of the six models to the neurobehavioral performance data of scenario 2 (both conditions). However, all data points immediately following awakening were removed prior to scaling of the model predictions and assessment of goodness-of-fit. Thus, the MSE and RRMSE values in this table were computed after specifically excluding data points likely to be influenced by sleep inertia. For comparison, the BMSE (estimated best possible MSE) was 40.92 and the WMSE (worst possible MSE) was 57.51 for this adjusted data set. Note that the updated predictions from the revised model E were used to compute that model's MSE and RRMSE values in this table.

The predictions provided for model A were considered separately, because they were dissimilar in various regards. Firstly, the entire set of predictions was shifted by 1 h with respect to the scenario schedules (possibly because the authors of model A used the wrong start time for the scenario). Secondly, the model software dictated the timing of the in-flight rest periods, which resulted in sleep/wake schedules deviating from the original scenario. Thirdly, the model distinguished between subjective sleepiness and neurobehavioral performance capability (although the predictions for these variables were fixed non-linear transformations of each other). The model A predictions for scenario 4 are shown in **Fig. 16**. The same scaling approach was applied as in Fig. 15.

Model to Data Comparisons for Scenario 5: Sleep Restriction and Recovery

Scenario 5 involved data from a laboratory experiment in which subjects were partially sleep deprived for 7 d; there were two experimental conditions. After the 7-d sleep restriction period, subjects received three days of recovery. This scenario is similar to scenario 2, but various details are different. Scenario 5 was first presented to the modeling teams at the Fatigue and Performance Modeling Workshop. Model predictions were requested by the end of the first day of the Workshop for overnight comparison to the experimental data. Model E was excluded from the model to data comparisons for this scenario, because prior to the Fatigue and Performance Modeling Workshop this model was revised and optimized using experimental data for the scenario. All five remaining modeling teams were

TABLE IX. QUANTITATIVE COMPARISONS OF MODEL PREDICTIONS TO EXPERIMENTAL DATA FOR SCENARIO 3.

Model	MSE
B	0.51
C	0.84
D	0.56
E	0.52
F	0.50

This table shows the MSE (mean square error) values resulting from the statistical comparisons of the five models for which predictions were provided to the combined experimental data (4-point alertness scale) for the 10 engineers of scenario 3. The results are listed by model label (see Table I) in alphabetical order. For comparison, the WMSE (worst possible MSE) value was 0.88 for this scenario. The BMSE (best possible MSE) could not be computed due to the lack of a population average time series.

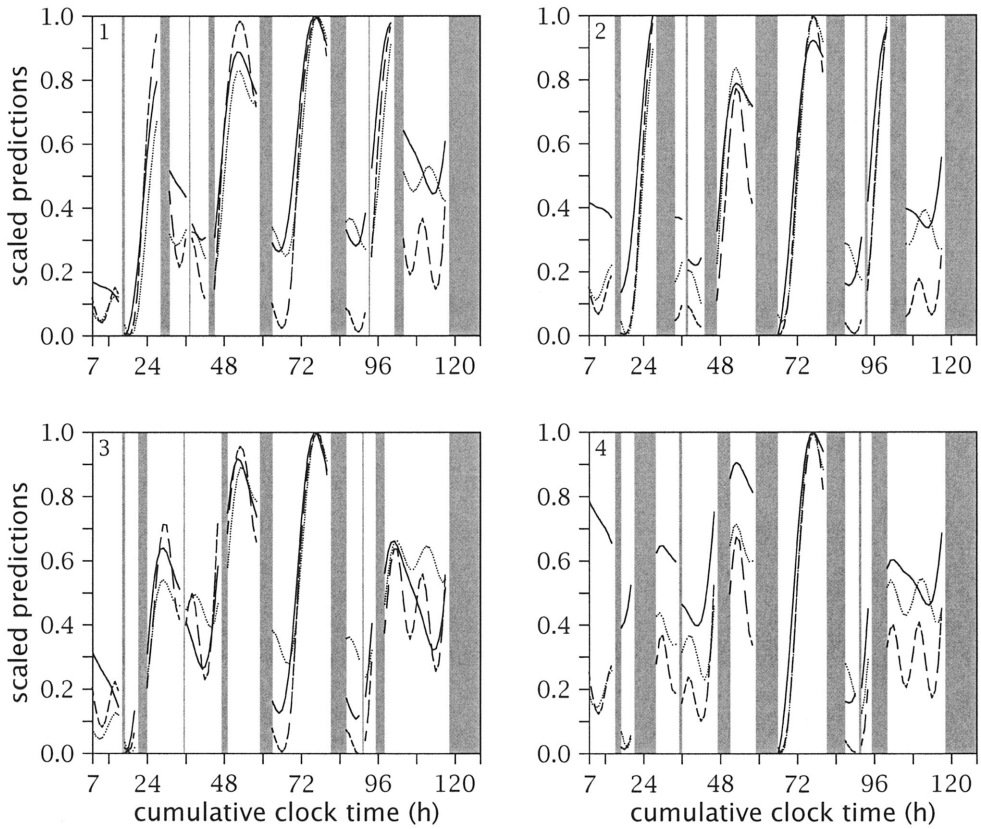

Fig. 15. Graphical comparison of subjective sleepiness predictions for scenario 4. Predictions from model B (dashed curves), E (dotted curves), and F (solid curves) are shown for pilots 1 (top left panel), 2 (top right panel), 3 (bottom left panel), and 4 (bottom right panel). Each model's predictions were linearly scaled so as to range from 0 to 1, with 0 and 1 corresponding to the lowest and highest sleepiness value, respectively, among the combined predictions for the 4 pilots. The abscissa shows cumulative clock time (in hours). Vertical gray bars indicate sleep opportunities. The differences among the pilots in each model's predictions for the first few hours of the scenario, before there is any variability in the pilots' schedules, may reflect differences in sleep/wake/work history.

able to provide predictions before midnight of the first Workshop day. However, modeling team D did not provide predictions for the 12:30 time points requested in the scenario description. In order to compare the models to the data on equal footing, therefore, the data for 12:30 were not used.

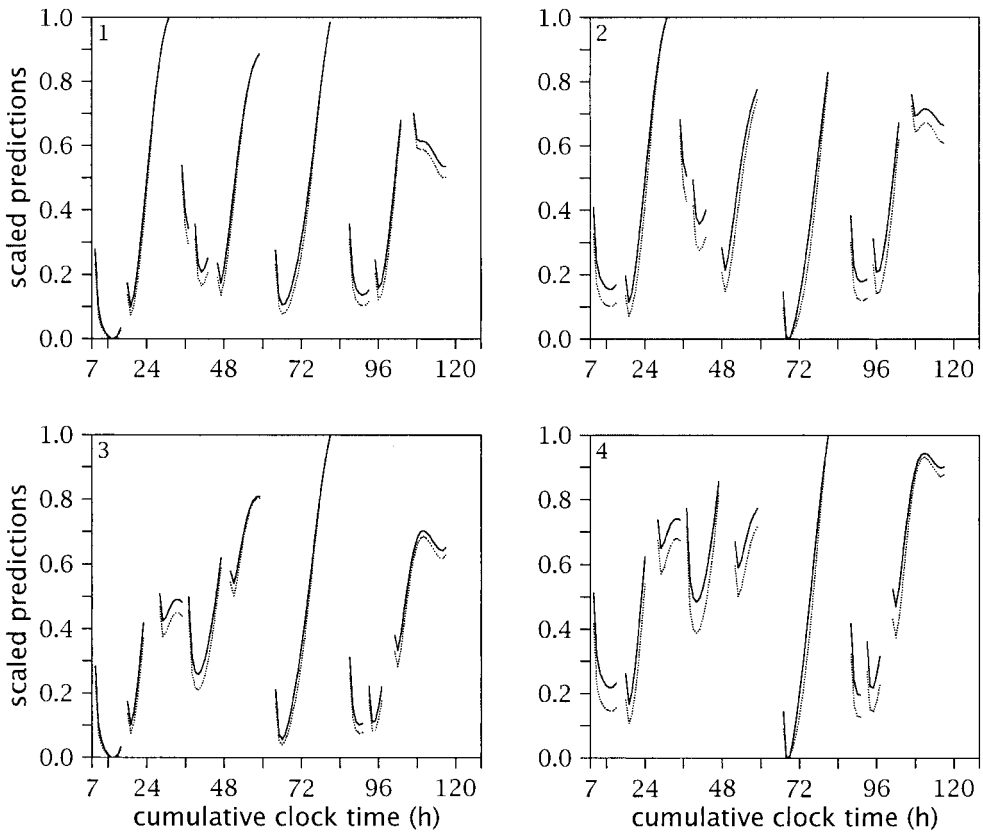

Fig. 16. Graphical representation of model A predictions for scenario 4. Subjective sleepiness predictions (solid curves) and neurobehavioral performance predictions (dotted curves) are shown for pilots 1 (top left panel), 2 (top right panel), 3 (bottom left panel), and 4 (bottom right panel). Both sets of predictions were linearly scaled so as to range from 0 to 1, with 0 and 1 corresponding to the lowest and highest value, respectively, among the model's combined predictions for the 4 pilots. Model A predictions deviated from the scenario in that the entire set of predictions was shifted by 1 h, and that in-flight rest periods were timed by the modeling software instead of the scenario schedule. The timing of other periods for sleep appeared to differ from the scenario schedule as well; gaps in the prediction curves indicate where sleep periods were placed (cf. Fig. 15). The abscissa shows cumulative clock time (in hours).

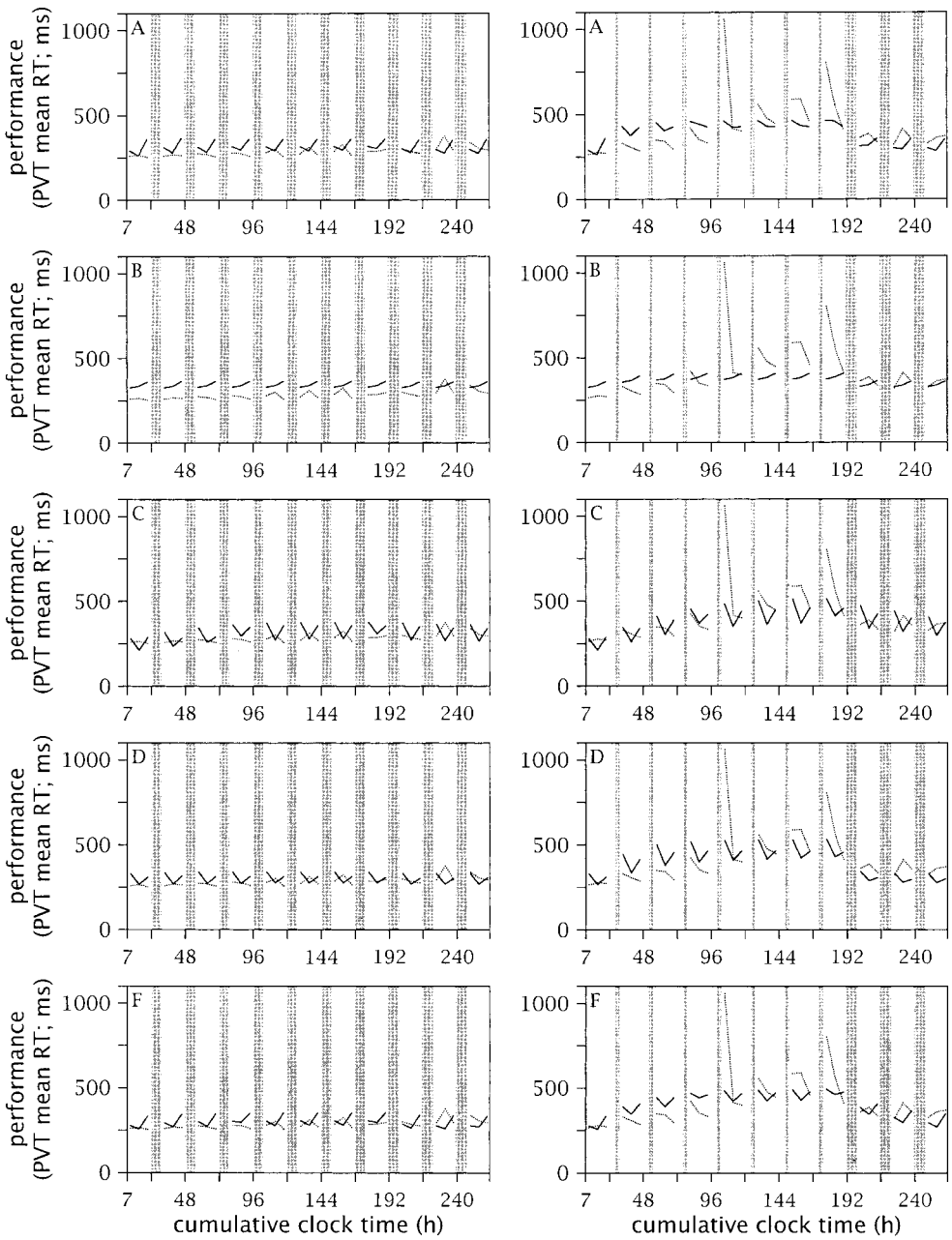

Fig. 17. Graphical comparison of models to neurobehavioral performance data for scenario 5. The left panels A through D and F show the scaled predictions (solid curves) from the corresponding models (see Table I), overlaid on the average experimental data (dotted curve), for condition 1. The right panels A through D and F show the same for condition 2. The model predictions were scaled to the data of both conditions simultaneously. The abscissa shows cumulative clock time (in hours) since awakening from the last baseline sleep period. The gray bars indicate sleep periods. Model E was not included in the model to data comparisons for scenario 5. Refer to the main text and to Fig. 6 for further details.

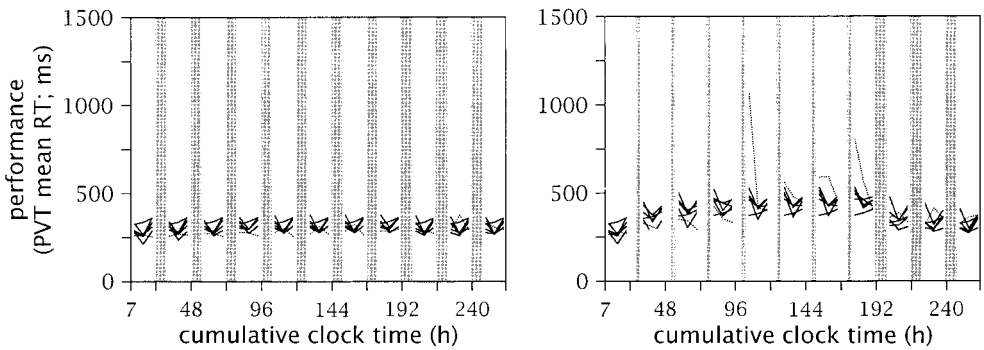

Fig. 18. Composite graphs for comparison of models to data in scenario 5. The graphs show the scaled predictions for all five models A through D and F (solid curves) overlaid on the average experimental data (dotted curve) in condition 1 (left panel) and condition 2 (right panel). Thus, the two panels combine the graphs in Fig. 17 by collapsing the panels in this figure from top to bottom.

TABLE X. QUANTITATIVE COMPARISONS OF MODEL PREDICTIONS TO EXPERIMENTAL DATA FOR SCENARIO 5.

Model	MSE	RRMSE
A	65472	82.39%
B	67580	96.97%
C	65718	84.10%
D	64066	72.53%
F	64844	78.00%

This table shows the MSE (mean square error) and RRMSE (relative root mean square error) values resulting from the statistical comparisons of model predictions to the experimental data for scenario 5. The data for neurobehavioral performance capability in this scenario were mean reaction times on a psychomotor vigilance task (PVT). The two conditions of the scenario were combined. The results of the model to data comparisons are listed by model label (see Table I) in alphabetical order. For comparison, the BMSE (estimated best possible MSE) was 54190 and the WMSE (worst possible MSE) was 68023. Model E did not participate in the comparisons for this scenario.

Fig. 17 shows the scaled model predictions x_t' overlaid on the average neurobehavioral performance data y_t' for each of the models. There were some "outliers" in the experimental data of condition 2, but these data points were not removed because they constituted actual observations with no known sources of systematic error. The linear mixed-effects regression method used to scale the model predictions to the data was relatively robust to these "outliers" and accordingly they did not affect the model to data comparisons much. Note that the scaling of the model predictions was performed for the two conditions in the scenario simultaneously.

Fig. 18 shows composite graphs with all five models overlaid on the average experimental data at once. **Table X** shows the MSE and RRMSE values resulting from the quantitative comparisons of the five models to the data for scenario 5, combining the data from the two conditions. As in scenario 2, it was found that the effects of chronic sleep restriction and recovery on waking neurobehavioral performance capability were difficult to predict by all of the models considered.

DISCUSSION

In pursuit of a key question of the Fatigue and Performance Modeling Workshop, "Where do current models converge?" (34), scenario 1 is perhaps the best to consider first. This scenario was based on a laboratory experiment in which subjects were either kept awake for 88 h straight (condition 1), or received 2-h nap opportunities every 12 h during 88 h of extended wakefulness (condition 2). The total sleep deprivation condition (condition 1) is illustrative of two facets that all models (listed in Table I) appeared to have as a common basis: fatigue increased and performance decreased progressively with prolonged wakefulness; and fatigue and performance varied over time in accordance with circadian rhythmicity. These facets are reminiscent of the two-process model of sleep regulation (2), which posits two primary regulatory components (10): a sleep/wake-related component that builds up across time of wakefulness and declines during sleep; and a circadian component that oscillates with (near-)24-h periodicity. In condition 1 of scenario 1, all model predictions—as well as the experimental data—reflected these two components. To facilitate discussion of comparisons among the models, the terms "sleep/wake-related component" and "circadian component" will be used to refer to these facets regardless of what they are called or how they are implemented by the respective modeling teams.

The shape of the sleep/wake-related and circadian components differed among models (Fig. 8–10, left panels). The build-up of the sleep/wake-related component over 88 h of sleep deprivation ranged from near-linear in models E and F to rapidly saturating in models B and D. The circadian component ranged from near-sinusoidal in models A and D to skewed in model F; models B and E showed prominent harmonic oscillations (i.e., 12-h or shorter periodicities) in the circadian component; and model C had a sawtooth appearance. Moreover, the relative contributions of the sleep/wake-related and circadian components varied among the models. For instance, the contribution of the circadian rhythm was relatively small in the model E predictions for condition 1. The experimental data for this condition displayed a greater relative contribution of circadian rhythmicity. Models B and C showed evidence of changes in the relative contribution of the circadian rhythm over time of sleep deprivation. Model B in particular thereby captured a dominant feature of the changes in subjective sleepiness over 88 h of total sleep deprivation.

The influence of these differences among the models is further exposed in condition 2 of scenario 1 (Fig. 8–10, right panels), where nap opportunities allowed the sleep/wake-related component to partially dissipate every 12 h. Immediately after each nap, however, there was a potential contribution from sleep inertia, which is the sleepiness and performance impairment commonly experienced right after awakening (15). Thus, in condition 2 there were two additional sources of differences among the models: the predicted rate of dissipation of the sleep/wake-related component during nap sleep, and the predicted magnitude of sleep inertia on awakening. Predictions of sleep inertia could also be seen after awakening from baseline sleep at the beginning of the 88-h period of wakefulness extension in both conditions, particularly for model A (see panels A in Figs. 8 and 9). Sleep inertia was not clearly present in the data at the beginning of the 88-h experimental period, but it was seen after awakening from the naps (38). The differential influences of dissipation of the sleep/wake-related component, sleep inertia, and circadian rhythmicity on the data of condition 2 are difficult to disentangle due to the 12-h regularity of the nap opportunities. Therefore, any further interpretation of differences among the models for their features within days in this condition would be mere speculation.

Table VI shows that for the two conditions of scenario 1 combined, the models did not differ substantially in their overall predictive capabilities. Model B ranked first for predicting the subjective sleepiness data, but not for predicting the psychomotor vigilance performance data. Similarly, model D ranked first for predicting performance, but not for predicting sleepiness. Dif-

ferent measures of fatigue and performance may respond differently to wakefulness extension, as has been suggested by the literature (11). Only the predictions from model D were considerably distinct for subjective sleepiness vs. performance outcomes, which may have contributed to this model's relatively good overall ranking for scenario 1. The consistency by which all models underestimated the data in condition 1 and overestimated the data in condition 2 or vice versa (depending on the measure) reveals that simultaneously predicting the temporal profiles for both conditions caused problems for all the models. This may be a result of misestimation for the dissipation of the sleep/wake-related component during the naps in condition 2.

While all the models predicted the fatigue and performance data collected under acute extension of wakefulness fairly well, none of them could predict the sleep dose-dependent build-up of impairment over multiple days of sleep restriction in scenarios 2 and 5 (Fig. 13 and 18). This may reflect the influence of the two-process model of sleep regulation (2,10), which to varying extents appears to have provided a basis for all six fatigue and performance models presently considered. Like these six models, the two-process model would predict saturation of day-to-day changes in alertness after just a few days of partial sleep deprivation, but this is not in agreement with experimental data (37,39).

A revised version of model E, optimized using data from scenario 5, predicted the build-up of performance deficits in scenario 2 much better (Fig. 14). If the other modeling teams would have had the opportunity to calibrate their models to the same data set, their predictions for scenario 2 also could have changed. Still, with the revised model E it remained a challenge to predict the temporal profiles of the data for both conditions of scenario 2 simultaneously. This suggests that the sleep dose-dependence of the rate of cumulative impairment in the data was not yet accurately captured by the revised model.

The recovery phase concluding the sleep restriction experiments of scenarios 2 and 5 was of interest because the recovery from performance deficits was incomplete—even after 3 d in scenario 5 (8). To varying degrees, this was reflected in the predictions yielded by models A, C, and F as well. Since the predictions for the prior days of sleep restriction did not match the data, however, it is hard to estimate how accurate these models were with regard to just the recovery phase.

As the predicted changes in the sleep/wake-related component across days of sleep restriction were relatively small in scenarios 2 and 5 for all models, these two scenarios offered good opportunities to consider the model's predictions within days. For scenario 5, only three data points were available per day to compare all the models (and model E was excluded from the model to data comparisons). Focusing primarily on scenario 2 with ten (condition 1) or nine (condition 2) data points per day, therefore, it turned out to be helpful to address sleep inertia first. After the first few days in condition 1, which involved sleep restriction to 4 h time in bed per day, subjective sleepiness and especially performance deficits were greatest immediately after awakening. Thus, sleep inertia was the dominant feature of the variations within days in the data for this condition. Models A, C, and D succeeded in describing this phenomenon to some extent, although not at the magnitude displayed in the experimental data. Model C showed the dynamic increase of the magnitude of sleep inertia over the first few days of sleep restriction.

It should be pointed out that the psychomotor vigilance performance data were collected closer to the time of scheduled awakening than the subjective sleepiness data, which affected the degree to which sleep inertia influenced these measures. This information was not a priori available to the modelers, who, therefore, could not be expected to estimate the magnitude of sleep inertia very accurately. Re-analysis of the model to data comparisons after excluding the data points likely to be influenced by sleep inertia (see above) revealed that the overall accuracy of the fatigue and performance predictions for scenario 2 was not much affected by whether sleep inertia was taken into account or not. Other aspects of the changes in fatigue and performance within days may carry more weight.

The data for scenario 2 showed evidence of the so-called "post-lunch dip" on some days (Fig. 2 and 3); this is the temporary increase in sleepiness and decrease in performance that may be observed during the afternoon (18). The post-lunch dip has been mathematically described as a (12-h) harmonic oscillation in the circadian rhythm (26). Models B and E predicted a post-lunch dip, but graphical comparison to the data suggests that both models may have placed the dip too early during the day. Quantitative assessments of this issue are beyond the scope of the present paper.

Considering the regular placement of sleep periods at 24-h intervals in both scenarios 2 and 5, it would seem that further sources of predictable variability within days included only the circadian component and the within-day build-up of the sleep/wake-related component. All models appeared to resemble each other in the shape of their predictions for within-day changes when disregarding differences with respect to sleep inertia and the post-lunch dip. They all tended to overestimate subjective sleepiness and performance impairment at the end of the day. This may indicate a need to model a (non-linear) interaction between the circadian and sleep/wake-related components (13), the existence of which is subject to debate (1,14).

It is clear that scenarios 2 and 5 were full of challenges beyond the current state of knowledge about the neurobiology of fatigue and performance. The outcome was that none of the models predicted these chronic sleep restriction scenarios well. The predictive potential of model B for psychomotor vigilance performance in scenario 2 was almost the same as that of a horizontal line, while this model performed relatively well in scenarios 1 and 3. This illustrates the scope of the problem of predicting performance capability in situations of chronic sleep loss, which is hampered by the current limited understanding of this issue. New experimental research is needed to inform model developers about how to proceed with modeling fatigue and performance during chronic sleep restriction scenarios.

For all five scenarios considered at the Fatigue and Performance Modeling Workshop, sleep/wake schedules were provided, so that all the modeling teams would have available at least the minimum amount of information required to run their models. In many operational settings, however, no precise information about sleep times is available. In the transportation industry, for instance, vehicle operators' sleep times are often irregular and unknown—but work times are usually logged and, importantly, scheduled in advance. Therefore, models of fatigue and performance based solely on work times could be particularly useful in these settings. In scenario 3, model C was reportedly used in this fashion; that is, the model's predictions for this scenario were based solely on the information about work times (shown in Fig. 5). The results revealed that there is some merit to this approach (Table IX). Nevertheless, the sleep/wake-based predictions provided by the other modeling teams (except modeling team A, which did not provide predictions for scenario 3) described the experimental data more accurately. Thus, it seems that knowledge about sleep times contributes discernibly to the predictive potential of fatigue and performance models in operational settings.

Data from field experiments are inherently more noisy than data from most laboratory experiments, because of unknown and/or uncontrolled sources of variance in the field. An example is the use of caffeine as a countermeasure to sleepiness; caffeine is widely available and its use in the field is rarely controlled or monitored. When information about caffeine intake is available, however, model predictions could benefit from taking that information into account. Scenario 4 provided specific information about caffeine in a theoretical schedule for future ultra-long-range flight operations. None of the six models presently considered was capable of using this information. This exposes a gap in the state of the art of model development—the importance of dealing with countermeasures such as caffeine in biomathematical models of fatigue and performance is clear (7). Scenario 4 also contained information relevant to light exposure in this theoretical schedule. Light exposure influences alertness via documented mechanisms partly incorporated in model D (24). Modeling team D did not provide predictions for scenario 4, however, so that the effect of taking light exposure into account could not be evaluated.

The fatigue and performance predictions for scenario 4 (Fig. 15 and 16) differed among the four models for which predictions were provided. Given the lack of experimental data for this theoretical scenario, any interpretation of these differences is bound to be speculative. The future might reveal which aspects of the predictions for scenario 4 best reflect reality. There were some similarities among the models, however, which again point to a shared basis underlying them. In particular, there was overlap among the models for the predicted changes in fatigue and performance during the second day of the layover period. Like real human behavior, which displays long-term robustness to sleep/wake history, there is a tendency of all models combining a sleep/wake-related component and a circadian component to converge to a common stable profile regardless of prior states. There are other aspects of the models that distinguish them conceptually, though. These aspects may differentiate them more under certain scenarios not currently considered.

The present analyses did not statistically compare the models to each other directly for any of the five scenarios, but numerical differences among the models were generally small compared to the differences between the model predictions and the experimental data. Across the four scenarios for which data were available to evaluate the models, not one model clearly stood out as the overall best or worst. A limitation of the present model to data comparisons is that the statistical analyses focused on empirical and predicted data values but largely ignored the time dimension. Further analyses concerned with temporal relationships (e.g., cross-correlation) could provide additional information to compare the models to the data and to each other. It is also important to realize that models' ability to obtain a perfect fit to experimental data is limited by stochastic variability (neurobiological, experimental, or other noise) in the observations. This limitation should not have to include the variability associated with systematic inter-individual differences in fatigue and performance, however. Powerful mathematical tools are available nowadays to deal with stable inter-individual differences (29,36).

Taken together, the model to data comparisons revealed that the models were capable of predicting the data of scenarios 1 and 3 fairly well. Considering the challenging nature of these scenarios, this is an accomplishment that should not be underestimated. Due to the complexities of the neurobiology of fatigue and performance, it was not expected that the models would perform well for all of the scenarios. Indeed, exposing "What is missing from the current models" was one of the key aims of the Fatigue and Performance Modeling Workshop (34). The chronic sleep restriction scenarios caused significant problems for all the models. Given the relevance of chronic sleep loss in many operational settings, this may be an area deserving priority for further model development.

In conclusion, this paper comprised a snapshot characterization of where current fatigue and performance models converge and what is missing from them still. A series of commentaries in this journal issue is concerned with further discussing the strengths and weaknesses of each of the models. Much has already been achieved for modeling changes in fatigue and performance over time. Model development also continues, and today's state-of-the-art models are tomorrow's previous versions. Indeed, the present results suggest that substantial additional development is necessary to create reliable tools for prospective prediction of fatigue and performance across a broad range of circumstances. Important capabilities that will need to be added to current models involve the effects of chronic sleep loss, countermeasure use (e.g., caffeine intake), and light exposure (although this is partly addressed in model D), as well as inter-individual differences. Iterative modifications of the currently existing models may not

suffice to deal with these issues, however; a paradigm shift in the approach to modeling (29,31) may eventually be more successful. New experimental data will need to be acquired to elucidate this matter, and to provide information for the development of future fatigue and performance models.

ACKNOWLEDGMENTS

The author wishes to thank the individuals and their collaborators who graciously provided the scenarios and experimental data sets used in this paper: David Dinges (University of Pennsylvania School of Medicine) for scenarios 1 and 2; Steven Popkin (U.S. Department of Transportation Volpe Center) for scenario 3; Melissa Mallis (NASA Ames Research Center) for scenario 4; and Thomas Balkin (Walter Reed Army Institute of Research) for scenario 5. The author also wishes to thank each of the modeling teams for providing their model predictions; and David Neri (U.S. Office of Naval Research) and Roy Vigneulle (Anteon Corporation) for tireless interfacing with the modeling teams, and for helpful suggestions with regard to this paper. Two anonymous reviewers also provided valuable comments on the paper. The author is grateful to Jaques Reifman (U.S. Army Medical Research and Materiel Command), Greg Maislin (Biomedical Statistical Consulting), and Erik Olofsen (Leiden University Medical Center) for discussions of the statistical methodology. This work would not have been possible without the support of the steering committee for the Fatigue and Performance Modeling Workshop, and sponsoring of the Workshop by the U.S. Department of Defense, U.S. Army Medical Research and Materiel Command, Office of Naval Research, National Aeronautics and Space Administration, Air Force Office of Scientific Research, and U.S. Department of Transportation. This work was supported in part by NIH grants NR04281 and HL70154, AFOSR grants F49620-95-1-0388 and F49620-00-1-0266, NASA Headquarters grant NAG9-1161, NASA cooperative agreement NCC 9–58 with the National Space Biomedical Research Institute, and NASA cooperative agreements NCC 2-1077 and NCC 2-1394 with the Institute for Experimental Psychiatry Research Foundation.

APPENDIX A: EXPLAINED VARIANCE

The development of a biomathematical model of fatigue and performance typically involves postulating relevant mathematical equations, and fitting these equations to experimental data in order to estimate the model parameters. The various procedures used for model fitting (e.g., least squares, maximum likelihood) result in approximate maximization of the variance that the model has in common with the data and minimization of the remaining error variance. As such, the variance contained in the model can range from zero (when the model is a horizontal line) up to the variance in the data (when the model describes the data perfectly and no error variance remains). The ratio of the variance in the model predictions $\hat{y}$ to the variance in the experimental data y is called "explained variance" and is computed as:

$$R^2 = \mathrm{Var}(\hat{y})/\mathrm{Var}(y),$$

with

$$\mathrm{Var}(\hat{y}) = \sum_{i=1}^{m} [\hat{y}_i - \mathrm{Avg}(\hat{y})]^2/m,$$

where m is the number of available data points. The above equation for $\mathrm{Var}(\hat{y})$ is also used (*mutatis mutandis*) to assess Var(y). The R^2 statistic is frequently applied as a measure of goodness-of-fit, with greater R^2 values taken as evidence of a better fit of the model to the data ($0 \leq R^2 \leq 1$).

The dependence of R^2 on the variance in the model makes the interpretation of this statistic problematic. This is easily demonstrated by considering a data set consisting only of noise, for which the best model would be a horizontal line. The variance contained in that model is zero, resulting in $R^2 = 0$. Thus, on the basis of explained variance it would seem that the horizontal line is the worst possible model rather than the best, and an alternative model containing greater variance could erroneously be judged to provide a better fit. It is the dependence of the R^2 statistic on both the modeled signal and the remaining error variance in the data, actually, that makes its interpretation ambiguous. This dual dependence is unnecessary, and can be avoided by focusing only on the error variance. As such, the sum of the squares of the errors (i.e., deviations between the model and the data), from which the mean square error (MSE) used in the present paper is derived, provides an unambiguous measure of goodness-of-fit.

In truly prospective model validation, when predictions are made on a pre-determined metric and no fitting or scaling is necessary to compare the model with the data or with other models, the R^2 statistic is a flawed measure of goodness-of-fit altogether. In particular, gross overestimation of goodness-of-fit may occur when the model predictions contain much variance, but the model does not accurately capture the profile of the data (in fact, the value of R^2 may even exceed 1). Furthermore, when the prospective predictions of two alternative models are compared on the basis of explained variance, the model with the greatest variance will be favored regardless of how well it matches the data, since the covariance between model and data is not explicitly included in the computation of the R^2 statistic. In the case of model fitting only, the covariance is rather accounted for implicitly; for least-squares fitting, it can be easily shown that $\mathrm{Var}(\hat{y}) = \mathrm{Covar}(\hat{y},y)$ so that

$$R^2 = \mathrm{Var}(\hat{y})/\mathrm{Var}(y) = \mathrm{Covar}(\hat{y},y)/\mathrm{Var}(y).$$

The equivalence of model variance and covariance is established in the model fitting process, and is lost outside the context of fitting. For truly prospective modeling, therefore, the explained variance is not interpretable as a measure of goodness-of-fit. To illustrate the scope of this issue, consider that the R^2 statistic would not even distinguish between two models of circadian rhythmicity with identical amplitudes and shapes but opposite circadian phase positions, while these models would yield entirely different prospective model predictions for circadian troughs in performance. Thus, the use of explained variance as a generic tool for evaluating and comparing goodness-of-fit is not advisable. By focusing instead on the error variance (e.g., by using MSE), proper assessments of goodness-of-fit can be made regardless of context.

REFERENCES

1. Achermann P. Technical note: A problem with identifying nonlinear interactions of circadian and homeostatic processes. J Biol Rhythms 1999; 14:602–3.
2. Achermann P. The two-process model of sleep regulation revisited. Aviat Space Environ Med 2004; 75(3, Suppl.):A00–00.
3. Achermann P, Werth E, Dijk DJ, Borbély AA. Time course of sleep inertia after nighttime and daytime sleep episodes. Archiv Ital Biol 1995; 134:109–19.
4. Åkerstedt T, Folkard S, Portin C. Predictions from the three-process model of alertness. Aviat Space Environ Med 2004; 75(3, Suppl.):A00–00.
5. Åkerstedt T, Gillberg M. Subjective and objective sleepiness in the active individual. Intern J Neurosci 1990; 52:29–37.
6. Balkin T, Thorne D, Sing H, et al. Effects of sleep schedules on commercial driver performance. Washington, DC: U.S. Department of Transportation, Federal Motor Carrier Safety Administration; 2000; Technical Report: DOT-MC-00–133.
7. Balkin TJ, Kamimori GH, Redmond DP, et al. On the importance of countermeasures in sleep and performance models. Aviat Space Environ Med 2004; 75(3, Suppl.):A00–00.
8. Belenky G, Wesensten NJ, Thorne DR, et al. Pattern of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. J Sleep Res 2003; 12:1–12.
9. Belyavin AJ, Spencer MB. Modeling performance and alertness: the QinetiQ approach. Aviat Space Environ Med 2004; 75(3, Suppl.):A00–00.
10. Borbély AA, Achermann P. Sleep homeostasis and models of sleep regulation. J Biol Rhythms 1999; 14:557–68.
11. Broadbent DE. Performance and it measurement. Br J Clin Pharmacol 1984; 18:5S–11S.
12. Dawson D, Fletcher A. A quantitative model of work-related fatigue: Background and definition. Ergonomics 2001; 44:144–63.
13. Dijk DJ, Duffy JF, Czeisler CA. Circadian and sleep/wake depen-

dent aspects of subjective alertness and cognitive performance. J Sleep Res 1992; 1:112–7.
14. Dijk DJ, Jewett ME, Czeisler CA, Kronauer RE. Reply to technical note: Nonlinear interaction between circadian and homeostatic processes: Models or metrics? J Biol Rhythms 1999; 14:604–5.
15. Dinges DF. Are you awake? Cognitive performance and reverie during the hypnopompic state. In: Bootzin RR, Kihlstrom JF, Schacter DL, eds. Sleep and cognition. Washington, DC: American Psychological Association, 1990; 159–75.
16. Dinges DF, Powell JW. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. Behav Res Methods Instr Comp 1985; 17:652–5.
17. Folkard S, Åkerstedt T. A three process model of the regulation of alertness and sleepiness. In: Ogilvie R, Broughton R, eds. Sleep, arousal and performance: Problems and promises. Boston, MA: Birkhäuser, 1991; 11–26.
18. Hildebrandt G, Rohmert W, Rutenfranz J. 12 and 24 h rhythms in error frequency of locomotive drivers and the influence of tiredness. Intern J Chronobiol 1974; 2:175–80.
19. Hursh SR. Modeling sleep and performance within the integrated unit simulation system (IUSS). Natick, MA: Natick Research, Development and Engineering Center; 1998; Technical Report: Natick/TR-98/026L.
20. Hursh SR, Redmond DP, Johnson ML, et al. Fatigue models for applied research in warfighting. Aviat Space Environ Med 2004; 75(3, Suppl.):A44–53.
21. Jewett ME, Forger DB, Kronauer RE. Revised limit cycle oscillator model of human circadian pacemaker. J Biol Rhythms 1999; 14:493–9.
22. Jewett ME, Kronauer R. Interactive mathematical models of subjective alertness and cognitive throughput in humans. J Biol Rhythms 1999; 14:588–97.
23. Jewett ME, Wyatt JK, Ritz-De Cecco A et al. Time course of sleep inertia dissipation in human performance and alertness. J Sleep Res 1999; 8:1–8.
24. Kronauer RE, Forger DB, Jewett ME. Quantifying human circadian pacemaker response to brief, extended, and repeated light stimuli over the phototopic range. J Biol Rhythms 1999; 14:500–15.
25. Mallis MM, Mejdal S, Nguyen TT, Dinges DF. Summary of the key features of seven biomathematical models of human fatigue and performance. Aviat Space Environ Med 2004; 75(3, Suppl.):A4–14.
26. Monk TH, Buysse DJ, Reynolds III CF, Kupfer DJ. Circadian determinants of the postlunch dip in performance. Chronobiol Int 1996; 13:123–33.
27. Moore-Ede M, Heitmann A, Croke D, et al. Circadian alertness simulator for fatigue risk assessment in transportation: Application to reduce frequency and severity of truck accidents. Aviat Space Environ Med 2004; 75(3, Suppl.):A107–18.
28. Moore-Ede MC, Mitchell RE. Method for predicting alertness and biocompatibility of work schedule of an individual. Lexington, MA: Circadian Technologies; 1995; Technical Report: United States patent # 5,433,223.
29. Olofsen E, Dinges DF, Van Dongen HPA. Nonlinear mixed-effects modeling: Individualization and prediction. Aviat Space Environ Med 2004; 75(3, Suppl.):A134–40.
30. Pollard JK. Locomotive engineer's activity diary. Washington, DC: U. S. Department of Transportation, Federal Railroad Administration; 1996; Technical Report: DOT/FRA/RPP-9601, DOT-VNTSC-FRA-96–12.
31. Reifman J. Alternative methods for modeling fatigue and performance. Aviat Space Environ Med 2004; 75(3, Suppl.):A173–80.
32. Roach GD, Fletcher A, Dawson D. A model to predict work-related fatigue based on hours of work. Aviat Space Environ Med 2004; 75(3, Suppl.):A61–9.
33. Spencer MB. The influence of irregularity of rest and activity on performance: a model based on time since sleep and time of day. Ergonomics 1987; 30:1275–86.
34. U.S. Army Medical Research and Materiel Command. Fatigue and Performance Modeling Workshop. Announcement brochure. Fort Detrick, MD, June 2002.
35. Van Dongen HPA, Dinges DF. Modeling the effects of sleep debt: On the relevance of inter-individual differences. Sleep Res Soc Bull 2001; 7:69–72.
36. Van Dongen HPA, Maislin G, Dinges DF. Dealing with inter-individual differences in the temporal dynamics of fatigue and performance. Aviat Space Environ Med 2004; 75(3, Suppl.): A147–54.
37. Van Dongen HPA, Maislin G, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep 2003; 26:117–26.
38. Van Dongen HPA, Price NJ, Mullington JM, et al. Caffeine eliminates sleep inertia: Evidence for the role of adenosine. Sleep 2001; 24:813–9.
39. Van Dongen HPA, Rogers NL, Dinges DF. Sleep debt: Theoretical and empirical issues. Sleep Biol Rhythms 2003; 1:5–13.
40. Verbeke G, Molenberghs G, eds. Linear mixed models in practice: A SAS-oriented approach. New York, NY: Springer; 1997.