scientific reports

Check for updates

OPEN

Dynamic ensemble prediction of cognitive performance in spaceflight

Danni Tu[1,7], Mathias Basner[2,7], Michael G. Smith[2], E. Spencer Williams[3], Valerie E. Ryder[3], Amelia A. Romoser[4], Adrian Ecker[2], Daniel Aeschbach[5,6], Alexander C. Stahn[2], Christopher W. Jones[2], Kia Howard[2], Marc Kaizi-Lutu[2], David F. Dinges[2] & Haochang Shou[1]✉

During spaceflight, astronauts face a unique set of stressors, including microgravity, isolation, and confinement, as well as environmental and operational hazards. These factors can negatively impact sleep, alertness, and neurobehavioral performance, all of which are critical to mission success. In this paper, we predict neurobehavioral performance over the course of a 6-month mission aboard the International Space Station (ISS), using ISS environmental data as well as self-reported and cognitive data collected longitudinally from 24 astronauts. Neurobehavioral performance was repeatedly assessed via a 3-min Psychomotor Vigilance Test (PVT-B) that is highly sensitive to the effects of sleep deprivation. To relate PVT-B performance to time-varying and discordantly-measured environmental, operational, and psychological covariates, we propose an ensemble prediction model comprising of linear mixed effects, random forest, and functional concurrent models. An extensive cross-validation procedure reveals that this ensemble outperforms any one of its components alone. We also identify the most important predictors of PVT-B performance, which include an individual's previous PVT-B performance, reported fatigue and stress, and temperature and radiation dose. This method is broadly applicable to settings where the main goal is accurate, individualized prediction of human behavior involving a mixture of person-level traits and irregularly measured time series.

Space travel is a costly and hazardous endeavor. Astronauts are often faced with cognitively demanding tasks that require sustained attention, despite chronic sleep deprivation and disruptions to their circadian rhythms[1]. Human performance deteriorates without proper sleep, manifesting in slower reaction times and increased errors[2], heightening the risk of operational accidents[3]. Therefore, it is critical to anticipate changes in alertness and performance on a dynamic and individualized basis[4]. Vigilant attention is a construct typically assessed using reaction time and accuracy-based metrics in tasks requiring sustained attention. While environmental and psychological correlates of vigilant attention have been studied in healthy humans on Earth[5,6], highly trained and carefully selected astronauts are not necessarily represented in this population. Astronauts are also exposed to a unique set of conditions in space[7–9], including microgravity, extended confinement and isolation, radiation exposure, and other environmental and operational extremes. The collective impact of these challenges on psychological health and performance is inconclusive[10,11] and not yet fully understood[12,13].

The goal of this study was to dynamically predict vigilant attention, assessed with a brief 3-min version of the Psychomotor Vigilance Test (PVT-B)[14], as a function of astronauts' past performance, self-reported stress and fatigue, demographic and operational information, and variations in environmental variables (Table S1). The main challenge to predicting PVT-B performance is unraveling variation associated with the circadian rhythm, individual traits, psychological state, and the external environment[15,16]. Previously, PVT performance

[1]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, 219 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, USA. [2]Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, 423 Guardian Drive, Philadelphia, PA 19104, USA. [3]Toxicology and Environmental Chemistry, National Aeronautics and Space Administration, 2101 E NASA Pkwy, Houston, TX 77058, USA. [4]Center for Toxicology and Environmental Health LLC, 2000 Anders Ln, Kemah, TX 77565, USA. [5]Department of Sleep and Human Factors Research, Institute of Aerospace Medicine, German Aerospace Center, Linder Höhe, 51147 Cologne, Germany. [6]Institute of Experimental Epileptology and Cognition Research, Faculty of Medicine, University of Bonn, Building 076, Venusberg-Campus 1, 53127 Bonn, Germany. [7]These authors contributed equally: Danni Tu and Mathias Basner. ✉email: hshou@pennmedicine.upenn.edu

	(N = 24)
Sex, n (%)	
Male	19 (79.2)
Female	5 (20.8)
Age at dock, years	48.2 (4.78)
Prior days in space	53.5 (72.7)
Prior missions	1.29 (0.86)
Highest educational attainment, n (%)	
Master's	14 (58.3)
MD/PhD	10 (41.7)
Nationality/agency, n (%)	
USA/NASA	16 (66.7)
Non-USA/non-NASA	8 (33.3)
Average pre-flight overall performance score (OPS)	0.95 (0.02)
Number of in-flight RST observations	87.2 (18.8)

Table 1. Summary characteristics of the astronauts with reaction self-test (RST) data. Table values are mean (standard deviation) and count (percent) for continuous and categorical variables, respectively. Due to astronaut privacy concerns, marital status and number of children are not reported in this table.

was predicted via a two-process model[17], which incorporates a system of differential equations to describe homeostatic and circadian pressures governing sleep. While such models have expanded our understanding of sleep regulation and alertness, and have been greatly adapted[18–20], they are often deterministic and so preclude statistical comparisons; even models with person-level random effects[21] cannot typically accommodate a large number of covariates.

Statistical models offer a complementary approach to prediction, focusing on prediction accuracy and uncertainty estimation at the expense of only indirectly modelling physiological processes. Traditional methods for assessing the associations between PVT performance and sleep patterns have included correlation and ANOVA analyses[22,23], which allow for hypothesis testing but cannot make forecasts of later performance, adjust for the autocorrelation in repeated PVT measures over time, or accommodate the non-linear relationships between PVT performance and predictors[24]. Methods which have addressed these obstacles have mainly considered mixed-effect models[25] or an ensemble of mixed-effects and random forest models[26]. However, neither of these methods can explicitly model time-varying predictors whose effects themselves are time-varying, as in the case of circadian effects[27] or acclimation[28,29].

In this paper, we propose a 3-model ensemble prediction scheme consisting of a linear mixed effects model[30], a random forest model[31], and a functional concurrent model[32], the last of which allows us to estimate time-varying effects of each (potentially time-varying) predictor. We also incorporate predicted outcomes from a two-process model[18] as a covariate, with the aim of connecting biomathematical and statistical models commonly used to predict PVT performance. Our method extends the 2-model ensemble proposed by Cochrane and colleagues[26], though we employ a variant of forward-chaining cross-validation[33] to assess model performance. We demonstrate that the ensemble best predicts over the entire mission compared to any single component alone.

Material and methods

Participants and protocol. Reaction Self-Test (RST; see "Reaction self-test (RST)" section) data were collected from N = 24 astronauts (Table 1) over 19 International Space Station (ISS) mission increments between 2009 and 2014[34]. Astronauts spent an average of 160 (SD = 19) days, with a range of 123–192 days, on the ISS. Two versions of the RST were used: a morning version was taken after awakening from sleep, and an evening version prior to bed. Ahead of spaceflight, astronauts were scheduled to complete the RST twice per testing day (i.e., one morning RST and one evening RST per day) at 180, 120, 90, 60, and 30 days before launch and daily in the week before launch. Post-mission RST assessments were scheduled daily in the week after return to Earth as well as once at 30, 60, and 90 days after return. During the space mission, astronauts were instructed to complete the RST twice a day every 4 days, with extra sessions completed around extravehicular activities (EVAs) and sleep period shifts to accommodate spacecraft dockings. The total adherence rate of 78.9% across all RSTs (83.8% in-flight) exceeded the pre-determined project goal of 75% adherence. This resulted in a total of 2968 RST observations. The original study and this retrospective analysis were approved by the Institutional Review Boards of Johnson Space Center and the University of Pennsylvania (for data analysis); all research was performed in accordance with relevant regulations and guidelines. Participants provided written informed consent prior to study participation and re-consented for this retrospective analysis.

Reaction Self-Test (RST). The RST consists of a short survey (described below) followed by a computerized and brief (3-min) version of the Psychomotor Vigilance Test (PVT-B). The PVT is a validated measure of sustained attention based on reaction time (RT) to visual stimuli that occur at random inter-stimulus intervals[35]. Astronauts were instructed to monitor a box on the laptop screen and press the space bar once a millisecond

Total number of observations	(N = 2968)
Period, n (%)	
Pre-flight	506 (17.0)
In-flight	2109 (71.1)
Post-flight	353 (11.9)
Time of day, n (%)	
Morning	1568 (52.8)
Evening	1379 (46.5)
Other	21 (0.71)
Alertness	
LRM-50	− 33.0 (12.7)
Sleep	
Time in bed sleeping, hours	6.61 (1.30)
Time in bed not sleeping, hours	0.61 (0.77)
Self-report 11-point ratings	
Low workload (0–10)	4.47 (2.18)
Very stressed (0–10)	3.87 (2.01)
Poor sleep quality (0–10)	3.60 (1.87)
Medication use	
Caffeine, doses	2.05 (1.50)
Sleep aid flag, n (%)	131 (4.41)
Decongestant flag, n (%)	25 (0.84)
Antihistamine flag, n (%)	36 (1.21)
Pain medication flag, n (%)	143 (4.82)
Extravehicular activity (EVA)	
EVA today flag, n (%)	8 (0.27)
EVA tomorrow flag, n (%)	23 (0.77)

Table 2. Summary measures from the reaction self-test (RST) data, including pre- and post-flight observations. Table values are mean (standard deviation) and count (percent) for continuous and categorical variables, respectively.

counter appeared in the box and started incrementing. After that, the RT was displayed for 1 s and the next stimulus was presented after a random inter-stimulus interval of 2-5 s. Participants were instructed to react as quickly as possible without hitting the spacebar in the absence of a stimulus. The PVT-B has been recognized as a sensitive tool for detecting the effects of acute and chronic sleep deprivation and circadian misalignment, both of which are highly prevalent in spaceflight[1,36]. It has negligible aptitude and learning effects[37], and is ecologically relevant as sustained attention deficits and slow reaction times affect many real-world tasks, including the operation of a moving vehicle[2].

Astronauts were instructed to perform the RST in the morning after getting up and in the evening in the 2 h prior to bed, though the hour of the day varied (Figure S1). While the PVT-B portion is the same in both morning and evening versions, other portions differ slightly. The survey portion of the RST includes a sleep diary and 11-point Likert-type rating scales on tiredness, mental fatigue, physical exhaustion, stress, sleepiness, and a final rating depending on the time of day: workload (evening administration only) or sleep quality (morning administration only; Table S2). During both the morning and evening RSTs, astronauts were asked to list the name, dose unit, and doses taken of all medications ingested before going to bed the previous night (morning RST) and since awakening in the morning (evening RST). Additionally, in the evening RST, astronauts were asked to list caffeinated foods or beverages consumed since awakening in the morning (in both cases, "None" and "Decline to answer" were response alternatives). Astronauts were also asked whether they performed an EVA that day. This information was used to create binary variables for certain classes of medications and upcoming EVAs for each RST observation (Table 2).

Among the PVT-B performance metrics, we derived the LRM-50 as the outcome of interest, since it has been shown to be highly sensitive to sleep deprivation and has an approximately normal distribution[38]. LRM-50 is a likelihood ratio-based metric that is based on response time (RT) distributions derived from either a non-sleep deprived (non-SDP) state corresponding to the first 15 h of wakefulness, or a sleep deprived (SDP) state corresponding to hours 15 through 33 of wakefulness. In the original study, these distributions were derived from participants in a total sleep deprivation protocol who performed the PVT every 2 h[38]. The RT space was divided into 50 categories consisting of 49 RT intervals plus false starts. For a certain RT range, the likelihood ratio was calculated as the relative frequency of responses falling into the range under the SDP condition, divided by that of responses falling into the range under the non-SDP condition. Likelihood ratios greater than 1 indicate that responses in that range are more likely to be observed under the SDP condition compared to non-SDP, and conversely for likelihood ratios less than 1. The LRM-50 score is calculated by determining the RT range and

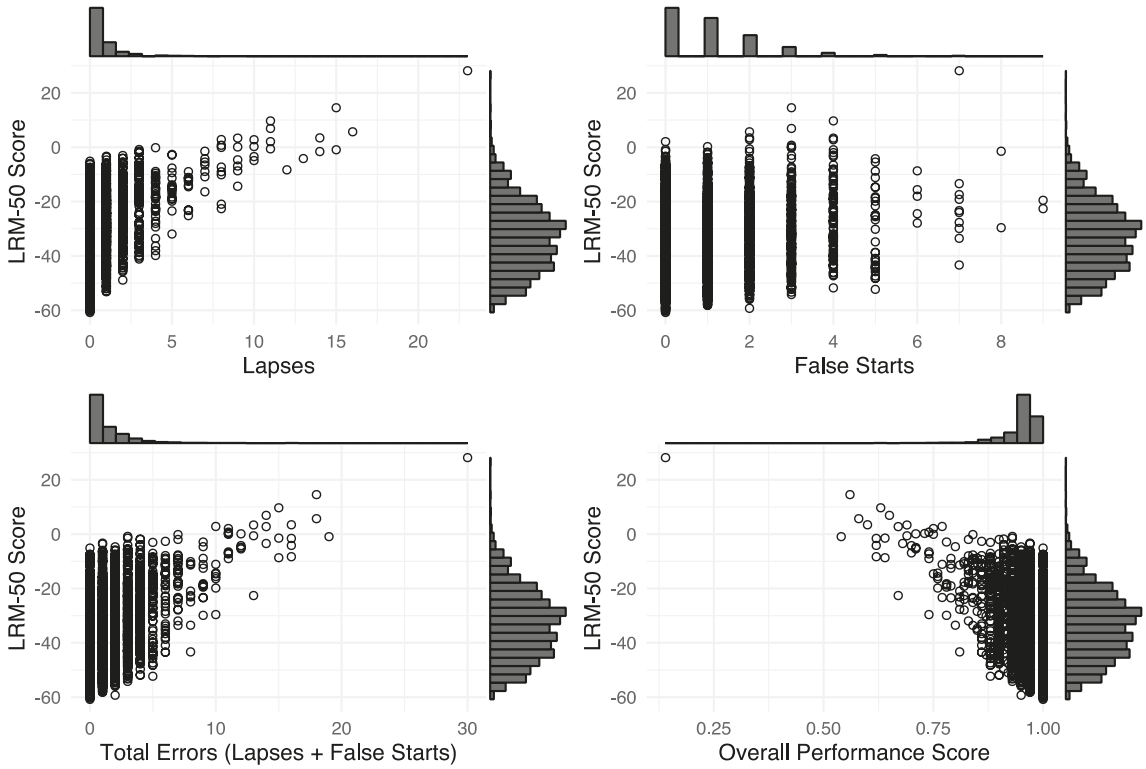


Figure 1. The LRM-50 score is best able to differentiate between the high performers in our sample, which includes pre-, post-, and in-flight observations (n = 2968) for all astronauts. These scatterplots show the joint distribution of LRM-50 with four popular PVT metrics: lapses, false starts, total errors (lapses + false starts), and the overall performance score (OPS). The histograms at the top and right edges show the marginal distributions of the variables on the x- and y-axes, respectively. Compared to other PVT metrics, the LRM-50 score is more normally distributed and is more sensitive to better performers (i.e., those with lower LRM-50).

associated likelihood ratio for each PVT-B stimulus. Likelihood ratios of all stimuli are then multiplied and log-transformed to induce symmetry around zero. Therefore, an LRM-50 of 0 means that this test bout is equally likely to be observed under an SDP or non-SDP condition. When LRM-50 < 0, the non-SDP condition is more likely relative to SDP, and conversely for LRM-50 > 0. LRM-50 correlates highly with response speed (reciprocal RT), but has the advantage that it also takes false starts (i.e., premature responses) into account. Compared to LRM-50, other commonly used PVT metrics[35], such as the number of lapses or false starts, were less effective in differentiating high performers such as astronauts (Fig. 1). We also considered a standardized measure of LRM-50, linearly scaled by each participant's mean and standard deviation.

Finally, as a person-level measure of baseline performance, we considered each participant's average pre-flight Overall Performance Score (OPS):

$$\text{Overall Performance Score} = 1 - \frac{\text{False Starts} + \text{Lapses}}{\text{Valid Stimuli}\ \left(\text{Including False Starts}\right)}.$$

The OPS is moderately sensitive to sleep loss, combines false starts and lapses into a single number, and is easily interpretable: an OPS of 1 corresponds to perfect performance, while 0 corresponds to the worst possible performance.

Environmental data. During the in-flight study period, five domains of environmental measures were recorded on the ISS. Radiation dose levels were obtained from the Space Radiation Analysis Group at NASA Johnson Space Center, and were summarized in daily absorbed dosage units (mGy) based on readings from dosimeters located aboard the ISS. The radiation dose was defined as the sum of radiation due to Galactic Cosmic Rays and the South Atlantic Anomaly. Measurements were collected from the following instruments over the course of the study: passive dosimeters (Radiation Area Monitor, 2009–2012), active dosimeters (Radiation Environment Monitor, 2012–2014), Tissue Equivalent Proportional Counter (TEPC, 2009–2012), and Intravehicular Tissue Equivalent Proportional Counter (IV-TEPC, 2013–2014). These instruments were rotated between several ISS modules (US Lab, Node 2, JEM, Columbus Module, and Service Module) at different times during the study period (Fig. S2).

Next, oxygen (O_2) and carbon dioxide (CO_2) levels in units of mmHg were collected from Major Constituent Analyzer (MCA) sample inlet ports located throughout the space station's air circulation system (Fig. S2, Panel B). Samples were drawn from these inlet ports in a cyclical fashion and were analyzed in two mass

spectrometer-based MCA units located in Node 3 and the US Lab. These units alternated between being the primary or backup unit, ensuring redundancy during scheduled maintenance or malfunction. We used the reading from whichever served as the primary sensor at the time. Temperature in Celsius (°C) was measured by sensors in the Node 2, Node 3, and US Lab modules. Temperature, CO_2, and O_2 data was downloaded from the Java Mission Evaluation Workstation System in intervals of at least 1 reading per second.

Noise exposure in A-weighted decibels (dBA) was not continuously monitored, but was rather collected periodically. Astronauts set up acoustic dosimeters at rotating locations aboard the ISS for 24-h periods approximately every other month. The internal memory of these dosimeters allowed the recording of dBA levels in one-minute intervals, which were acquired through manual display recall and infrared serial interface download.

Demographic and operational data. Demographic information was obtained from all astronauts (Table 1) and included sex, age at the time of ISS docking, nationality, space agency, educational attainment, number of prior space missions, and prior days in space. Operational data included the number of occupants on the ISS for each day of the mission and proximity of test to dock/undock maneuvers or EVAs.

Derived predictors. We derived several predictors based on the RST data: a stress/fatigue composite score, four medication flags, and predicted PVT lapses given the sleep schedule. The stress/fatigue composite score was created based on principal components analysis (PCA) of the 11-point scales on which the crew rated several behavioral states (i.e., sleepiness, tiredness, fatigue, exhaustion, stress, workload, and sleep quality) before taking the PVT. The score was calculated as the weighted average of the 11-point rating questions, with weights determined by the loadings onto the first principal component (PC). Table S3 shows the loadings onto the first PC, which accounted for 48.3% of the variance. Higher values of the stress/fatigue composite variable correspond to increased tiredness, more stress, and worse sleep quality. The loading for workload was negligible, potentially due to its lack of correlation with the other variables. Next, medication use was coded as a binary variable for four broad categories: pain medications, sleep aids, decongestants, and antihistamines. These categories were chosen due to their established use by astronauts on the ISS[39] and because their use may affect sleep or be correlated with conditions affecting sleep or alertness[40–42]. To represent the complex information contained in the sleep schedule, the final derived covariate was the number of predicted PVT lapses under a two-process model[18], which was calculated solely using an individual's reported bedtime and wake time.

Data integration and interpolation. To integrate environmental data with RST data, several strategies were required as different variables were recorded at different time intervals: RST was collected twice a day every 4 days; radiation dose and other operational variables were measured daily; temperature, noise level, CO_2, and O_2 were measured multiple times per day or minute (Table S4). For each RST observation, the value of radiation and ISS occupancy from that day was used. For temperature, noise, CO_2, and O_2, we used the average during the hour that the RST was completed, if available. Due to the logarithmic nature of decibel units, noise values were always averaged using the energetic average, while all other variables were averaged using the usual arithmetic mean.

During some 1-h periods, noise and temperature measurements were available for more than 1 sensor. For RST observations that occurred during these times, we employed location matching to achieve the best estimate for that individual. When the RST was taken on a computer located in Node 2 (where the crew quarters are located) or the US Lab, only the temperature or noise data from the corresponding Node 2 or US Lab was used. When the RST was taken elsewhere or the location was unknown, a weighted average of the Node 2 (75%) and US Lab (25%) measurements were used, reflecting the approximate empirical frequency of RSTs taken in these modules. Other variables (CO_2, O_2, and radiation) were only measured from one sensor at a time, so they were not location-matched.

When the daily or hourly value of an environmental variable was unavailable, we used two interpolation strategies, which are summarized in Table S4. Temperature, O_2, and CO_2 data had a relatively low rate of missingness (Table S5), so the locally estimated scatterplot smoothing (LOESS, neighborhood parameter $\alpha = 0.1$) value was used when the hourly average was not observed. Enough temperature data was available for Node 2 and US Lab that we could fit 3 separate LOESS interpolations: one for RST observations in Node 2, one for US Lab, and one for other or unknown locations which were interpolated using the weighted average described previously.

Noise levels were recorded on only 47 occasions throughout the study period, for a total of 722 hourly values. However, RST observations only rarely coincided with days that the acoustic dosimeters were active; as a result, 99.71% of in-flight RST observations did not have an observed noise level. For interpolation, we assumed that noise levels followed a 24-h cycle that was similar between days, with higher noise levels during the daytime (defined as 7:00 AM to 10:59 PM UTC) compared to nighttime (11:00 PM to 6:59 AM UTC) (Figure S3). Then, noise levels were interpolated using separate linear interpolations for averaged daytime and nighttime values. In other words, the noise level for a daytime RST observation was the interpolated value between the average daytime noise level from the most recent day of noise collection, and the next upcoming day; the same procedure was used for nighttime noise. The distribution of smoothed and unsmoothed environmental data is shown in Fig. 2.

Finally, the predicted PVT lapses depended solely on each individual's reported sleep schedule consisting of bedtimes and wake times. For RST observations where the sleep time was not reported, these variables were carried forward from the last observation for that individual, and the predicted lapses were calculated using the observed and interpolated bedtimes and wake times. One individual did not report any sleep data, so their predicted lapses were replaced by the overall average.

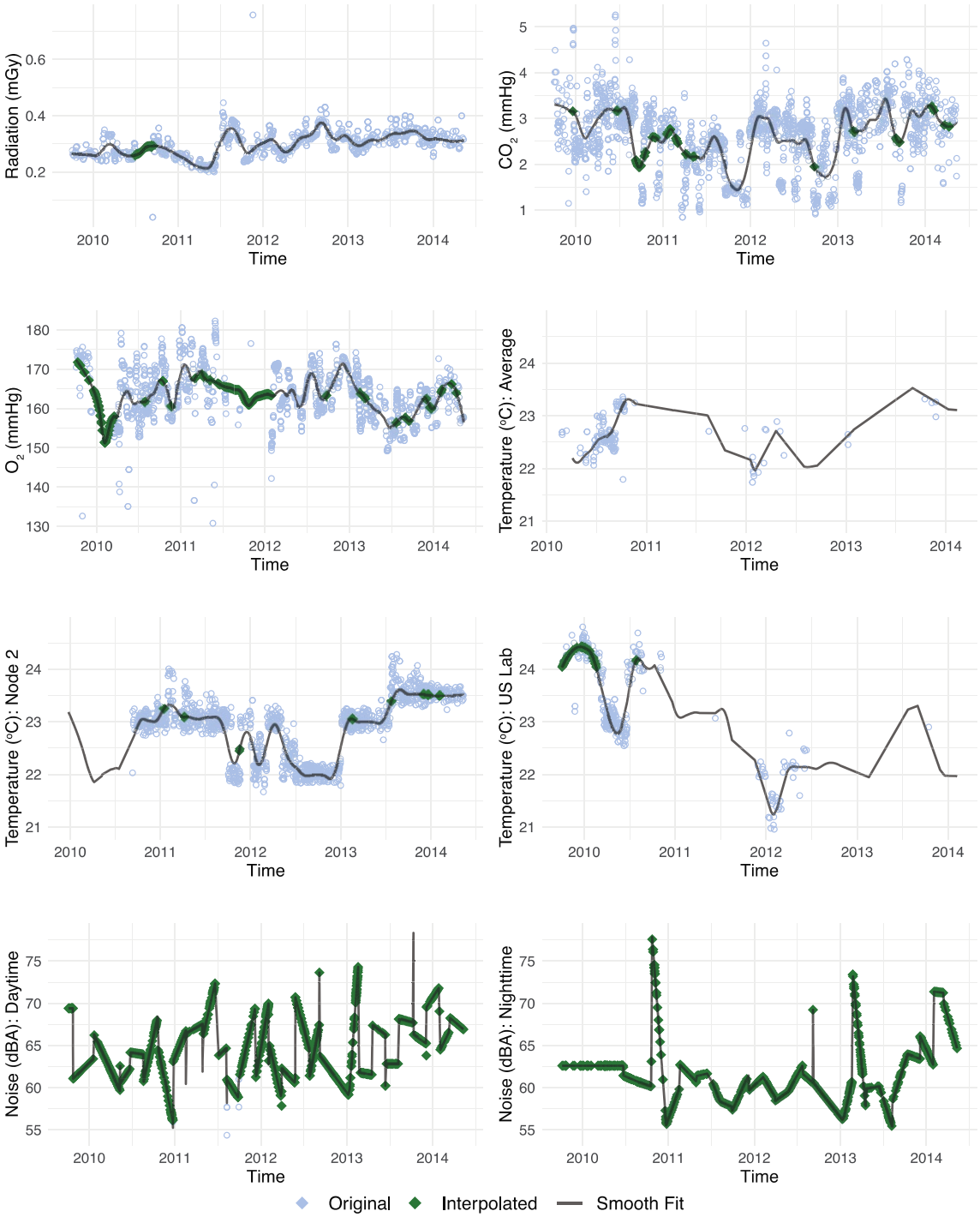


Figure 2. For each RST observation, the corresponding value of the environmental variable was found by using the observed value (if available) or the interpolated value formed by neighboring observations ("Data integration and interpolation" section). These plots illustrate the LOESS curves (black line) fit to the entire environmental data for radiation, temperature (separately for each location), CO_2, and O_2. A linear interpolation was used for noise (separately for daytime and nighttime). Each hollow blue circle corresponds to the observed hourly average (CO_2, O_2, temperature) or daily average (noise, radiation) that was used for an RST observation; the green diamond indicates that the interpolated value was used.

Statistical models. Our main goal was to construct a statistical model to predict the LRM-50 score for each participant at future points in time. Of the variables collected, we were also interested in identifying a subset of variables that were most important to predicting LRM-50. Candidate predictors of LRM-50 included a mixture of time-varying (i.e., function-valued) variables such as environmental data, most recent LRM-50 score, self-

reported stress/fatigue score, and ISS occupancy, as well as person-level (i.e., scalar-valued) data including each participant's demographics, pre-flight average PVT, sex, and age at docking.

We employed an ensemble of several models to address each aspect of the data. For participant i and time t, the linear mixed effects (LME) model defines the LRM-50 score y_{it} as a function of p covariates $X_{it} = \left(X_{it}^{(1)}, \ldots, X_{it}^{(p)}\right)$, intercept β_0, a p-dimensional vector of fixed effects β, person-specific random intercept b_i, and error ε_{it}:

$$y_{it} = \beta_0 + X_{it}\beta + b_i + \varepsilon_{it}.$$

The advantages of LME include its simplicity and efficiency, as well as the option to model correlated measurements over time: we specified a lag-one autoregressive (AR1) correlation structure to model the repeated measures of y_{it}.

By contrast, the random forest model[31] specifies no closed form for the relationship between y_{it} and X_{it}; rather, it uses an aggregate of decision trees to identify splitting points for continuous variables that optimally predict the outcome. While prone to overfitting, random forests are able to model a more flexible non-linear relationship between outcome and predictors, at the cost of interpretability.

Finally, since neither the random forest nor the LME are able to model the serial dependence of time-varying predictors and their time-varying effects, we also considered the functional concurrent model[32]: for participant i and observation j at time t_{ij}, the time-varying outcomes y_{ij} are related to p covariates $X_{ij}^{(1)}, \ldots, X_{ij}^{(p)}$ through the following:

$$y_{ij} = \beta_0 + f_1\left(X_{ij}^{(1)}, t_{ij}\right) + \ldots + f_p\left(X_{ij}^{(p)}, t_{ij}\right) + b_i(t_{ij}) + \varepsilon_{ij},$$

where f_i are smooth functions approximated by thin plate splines, and $b_i(t)$ and ε_{ij} are Gaussian processes representing person-level random trajectories and time-independent errors, respectively. The ensemble prediction was then constructed as the average of the predictions from the LME, random forest, and functional concurrent model. All data analyses were performed using R version 3.6.1[43], employing the *nlme*, *randomForest*, and *fcr* packages for each model.

Model validation. To assess the performance of each model as well as the ensemble, we employed a forward-chaining validation procedure (Fig. 3). For participant i with n_i RST observations, training length $t \in \{5, 10, \ldots, 45, 50\}$, and window number $k \in \{1, 2, \ldots, \min(20, n_i - t + 1)\}$, a given model was fit on window k defined by the k through $(k + t)$th RST observations from participant i and all RST observations from all other participants. That is, the model is trained on a partial time series (window k) for person i plus all other individuals' full time series. Then, the squared prediction error was assessed at both observation $k + t + 1$ (test) and window k (training). By incrementing t, the size of the training set is allowed to increase; by incrementing k, the training set shifts in time, so that the model is trained and evaluated at different portions of the mission. In contrast to methods which define the training set as *all* points prior to observation $t + k + 1$[26], controlling t allows us to minimize biases due to individual heterogeneity in observation frequency and number.

To aggregate these point-level errors into an overall error metric over all individuals, we defined a metric that weighted individuals equally, despite variation in n_i. For a given number of training days t, prediction error in the test set was measured at several levels. First, we defined the ith person's error by the average squared error at day $t + k + 1$ (i.e., the test set) over all windows k:

$$MSE(i, t) = \frac{1}{N_{i,t}} \sum_{k=1}^{N_{i,t}} \left(y_{itk} - \hat{y}_{itk}\right)^2,$$

where y_{itk} is the true LRM-50 at day $t + k + 1$, $\hat{y}_{itk}$ is the predicted value, and $N_{i,t} = \min(20, n_i - t + 1)$ is the number of possible window shifts. Then these were averaged over all $n = 24$ participants:

$$MSE(t) = \frac{1}{24} \sum_{i=1}^{24} MSE(i, t).$$

This averaging procedure ensured that, although some participants did not have enough data for 20 window shifts, their errors were weighted equally.

We then calculated the overall mean squared error (MSE) over all 10 values of t:

$$MSE_{overall} = \frac{1}{10} \sum_{t \in \{5, \ldots, 50\}} MSE(t).$$

As before, $MSE_{overall}$ weights errors equally for all participants and all values of t. To calculate errors for the *training* set, we followed the same procedure as for test errors $MSE(i, t)$, except we also summed over the t observations in the partial time series (window k) used in training the model:

$$MSE_{train}(i, t) = \frac{1}{N_{i,t}} \sum_{k=1}^{N_{i,t}} \sum_{j=1}^{t} \left(y_{itkj} - \hat{y}_{itkj}\right).$$

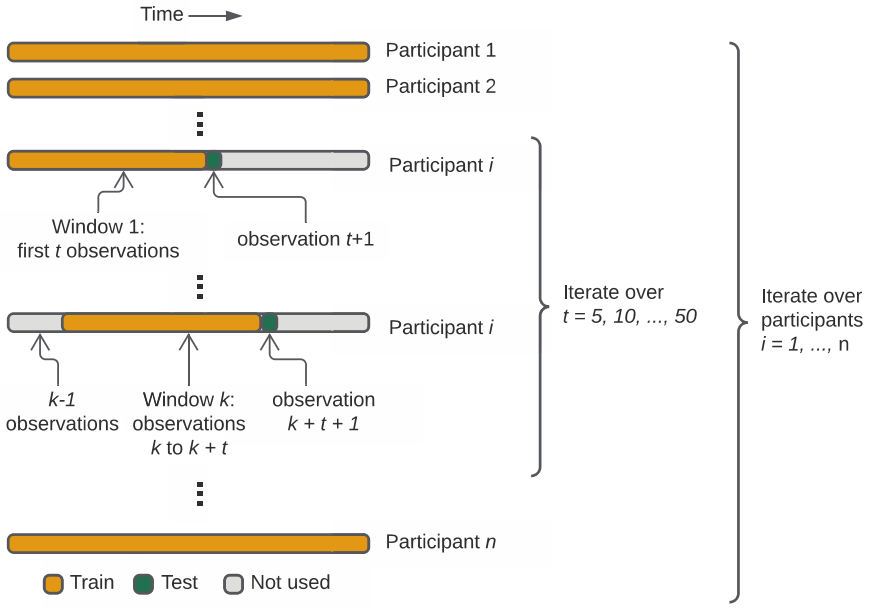


Figure 3. We used a forward-chaining procedure to assess prediction accuracy of a given model. For participant i and number of days t, the model was fit on window k defined by the k through $(k + t)$th Reaction Self-Test (RST) observations from participant i and the full data for all other participants. Then, the model prediction on the subsequent day was compared to the observed value on that day. The person-level prediction accuracy for a given participant was defined as the averaged squared difference between predicted and observed values over all values of k. The overall accuracy was then defined as the average of the person-level prediction accuracies.

Finally, to avoid over-penalizing large prediction errors, we considered an alternative error metric where the squared error $MSE(i, t)$ was replaced by the median absolute error (MAE):

$$\mathrm{MAE}(i,t) = \mathrm{median}\left\{\left|y_{itk} - \hat{y}_{itk}\right|\right\}_{k=1}^{N_{i,t}},$$

which is less sensitive to timepoints with large discrepancies between the predicted and observed LRM-50.

Other models, such as multivariate linear regression, time series regression (using the *dyn* R package), and generalized additive models (using the *gamlss* R package) were considered at this stage, but did not improve performance or goodness-of-fit of the final ensemble.

Variable selection. To identify the most important subset of variables for predicting LRM-50, we quantified a variable's importance by the average increase in MSE (%IncMSE) when permuting that variable within a random forest model. We also considered importance based on the increase in node purity, which is measured by the Gini index. This data-driven framework for feature selection has been previously deployed in behavioral contexts, including the identification of self-assessed and imaging biomarkers of cognitive impairment[44,45]. Through Monte Carlo sampling of 50% of the data, we obtained 100 rankings of variable performance. The most important variables were then defined as those that appeared in the top 10 with the highest frequency (Fig. S4). While this metric ("Top 10 Rate") identifies variables that are consistently important to prediction, those scoring lower are not necessarily uninformative. In statistical analyses, we considered both models fit on the full set of predictors, as well as the subset consisting of the most important variables.

Shiny application. The ensemble model was implemented as a user-friendly and interactive R Shiny application (Fig. 4). Given the data and the fitted ensemble model, the application displays individualized LRM-50 predictions in the context of their entire performance history, and other model diagnostic information. To encourage hypothesis generation, the value of predictor variables can also be "toggled," allowing the user to view how the predicted LRM-50 changes under hypothetical sets of conditions. Finally, the application includes each participant's entire trajectory of predicted and observed LRM-50 scores (not shown).

Results

Importance ranking of predictors of LRM-50. According to the random forest importance ranking, the most important predictors of LRM-50 were individual characteristics including age and average pre-flight OPS; the most recent LRM-50 score (lagged LRM-50); psychological factors including the composite stress/fatigue score; caffeine intake; total sleep missed during the most recent sleep opportunity (i.e., the sum of time taken to fall asleep, time spent awake during the night due to sleep disturbances, and time spent in bed before

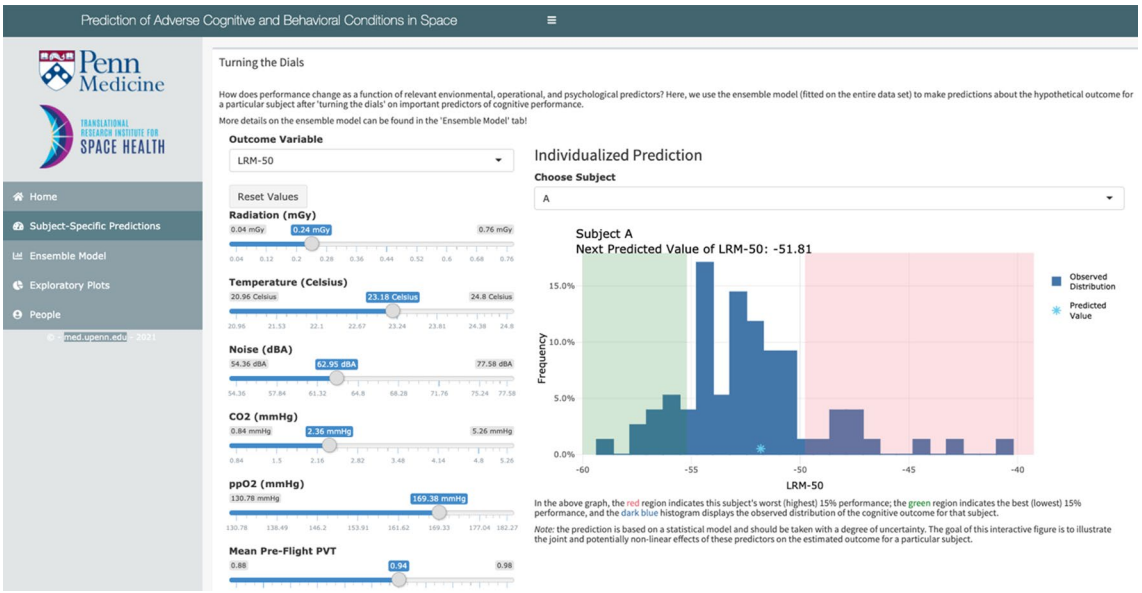


Figure 4. A screenshot of the R Shiny application implementing the ensemble prediction model. In the left panel, the user may "toggle" the value of each predictor (pre-set to averages observed for the individual astronaut). In the right panel, the resulting individualized predicted LRM-50 score for the selected participant is displayed (blue star at bottom of graph), along with the distribution of that astronaut's observed scores over the entire in-flight period. The prediction was made given the most up-to-date information for the astronaut, and the red and green regions correspond to that astronaut's worst 15% and best 15% scores overall.

getting up); and smoothed environmental measurements, including temperature, noise, and radiation dose (Fig. 5). Additionally, the first 10 variables (lagged LRM-50 through total sleep missed) had a relatively high Top 10 Rate compared to the other predictors, signifying that the rankings were stable from subsample to subsample.

Other sleep variables (sleep quality and sleep duration), O_2 and CO_2 levels, and sex were moderately important. The test track (morning or evening RST), medication use, ISS occupancy, scheduled EVAs, and workload were rated lower, meaning that they were not consistently among the 10 most important predictors of LRM-50. Variable rankings based on node purity (Fig. S5) were similar to those noted in Fig. 5, though node purity-based importance tended to be more stable across Monte Carlo iterations.

Prediction accuracy of the ensemble model. Our analyses indicate that the ensemble model performed better than any single model alone over various training lengths t after forward-chaining cross-validation (Fig. 6). In testing data, the ensemble model achieved the lowest MSE on average (Table S6); it out-performed all 3 of its constituent models for $t \geq 20$, but was out-performed by the LME for shorter training lengths $t < 20$. In the training data, the random forest model achieved the lowest MSE for all values of t.

Many of the largest MSE values were due to "spikes" in a participant's LRM-50 score that were unanticipated by all of the models (Fig. S6). In terms of the average MAE-based error, which is more robust to large prediction errors from spikes, the ensemble model continued to outperform the other models. However, its lead over the other models was less pronounced (Fig. S7).

While the previous results were based on the models' fit on the full set of covariates, we also asked if performance was preserved if the model was fit on a subset. This subset of "Top 10" variables was determined by the variable importance ranking in "Importance ranking of predictors of LRM-50" section: they were the lagged LRM-50, noise, temperature, stress/fatigue composite, pre-flight OPS, age, caffeine doses, radiation, "Very Stressed," and total sleep missed. We also considered the models' fit on the Top 10 variables excluding noise (due to its high interpolation rate) and including CO_2 and O_2 (which were environmental factors of interest and ranked highly in terms of node purity-based importance). This defined 4 additional models (Table S6), which all performed similarly, though models excluding the noise variable achieved the highest errors.

Modelling the outcome as a standardized LRM-50 score, scaled by each participant's average and standard deviation, led to similar performance as before, with the ensemble again outperforming its components (Fig. S8). Slight improvements in performance of the ensemble and functional regression models were achieved when replacing the LRM-50 score with the standardized version.

Finally, prediction accuracy was comparable when predicting more than 1 observation in the future. In addition to varying training length t, we also varied the length of the test set (i.e., the number of future observations to predict in each cross-validation loop), up to 7 observations. Because the RST was administered twice a day every 4 days, this corresponded to an average chronological horizon of 15.32 days ahead. Overall, prediction accuracy was stable as the horizon grew, with marginal increases in MSE for predictions further out (Table S7).

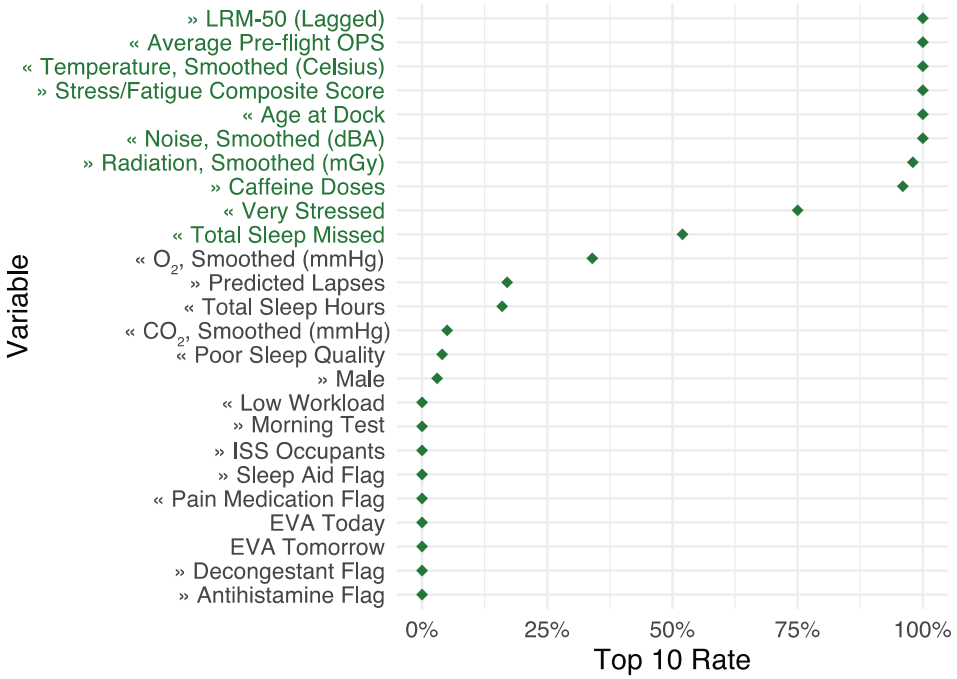


Figure 5. A variable's importance was measured by the increase in mean squared error (MSE) by permuting that variable in a random forest model. Variable importance rankings were obtained from 100 resampling draws. The resulting "Top 10 Rate" (x-axis) describes how a given variable, over resampling trials, is repeatedly among the 10 most important variables in a random forest model. We then defined the Top 10 variables as those which scored higher on this metric; these consisted of the lagged LRM-50 through Total Sleep Missed (green text). Arrows next to names refer to the direction of association in the linear mixed effects model (Table S8): "< <" represents association with lower LRM-50 (better performance) and vice versa for "> >". Due to low counts, the EVA variables were not included in the linear mixed effects model. (OPS = Overall Performance Score; RST = Reaction Self-Test; ISS = International Space Station; EVA = extravehicular activity).

Nonlinear and time-varying associations between environmental factors and alertness. To explore the modeled relationships between environmental conditions and LRM-50, we examined the corresponding LME coefficients (Table S8) and functional concurrent regression heat maps (Fig. 7). In both models, we found that better predicted outcomes (i.e., lower LRM-50) was associated with lower radiation dose, higher CO_2 exposure levels, and fewer ISS occupants. Noise levels appeared to be associated with better performance; however, the majority of noise observations were imputed, and we refrain from interpreting this finding.

While the LME estimated positive effects of increased O_2 and temperature on predicted LRM-50, the functional model revealed that these associations may be non-monotonic, suggesting that values either higher or lower than a certain range can affect vigilant attention negatively. In particular, partial pressures of O_2 between 160 and 170 mmHg were linked to better predicted performance. The model predictions also suggest better PVT-B performance at low partial pressures of O_2 (< 140 mmHg). Temperatures between 21.5 and 22.5 °C and higher than 23 °C were associated with better predicted performance.

In addition to the concurrent radiation levels, we calculated cumulative radiation doses for each person on each day of their mission, defined as the cumulative sum of daily exposure values in mGy. While cumulative radiation dose was found to be an important predictor for LRM-50 on top of concurrent radiation (Supplementary Methods 1), its inclusion did not ultimately improve model performance even when limited to variables ranked high in importance (Table S9).

Individualized predictions. An astronaut's LRM-50 score can be predicted at an arbitrary number of future time points, but this requires knowledge of environmental conditions and other covariates at those time points. In practice, we may obtain the best prediction at a particular time point by re-fitting the model on all previous data from that individual, as well as all data collected from other participants. After fitting the model, predictions are then made using the observed covariates from that day. By repeatedly re-fitting the model and predicting the next LRM-50 score at each observation, we are able to compare the entire sequence of the observed and predicted performance for each participant (Fig. 8). To estimate sampling variation, bootstrap confidence intervals and interquartile range can be obtained by bootstrapping participants (i.e., sampling individuals' entire time series with replacement). We bootstrapped at the level of participants in order to preserve trends across the mission. The root mean squared error (RMSE) of these "chained" predictions over time ranged from 3.93 to 12.11 among astronauts (Fig. S9).

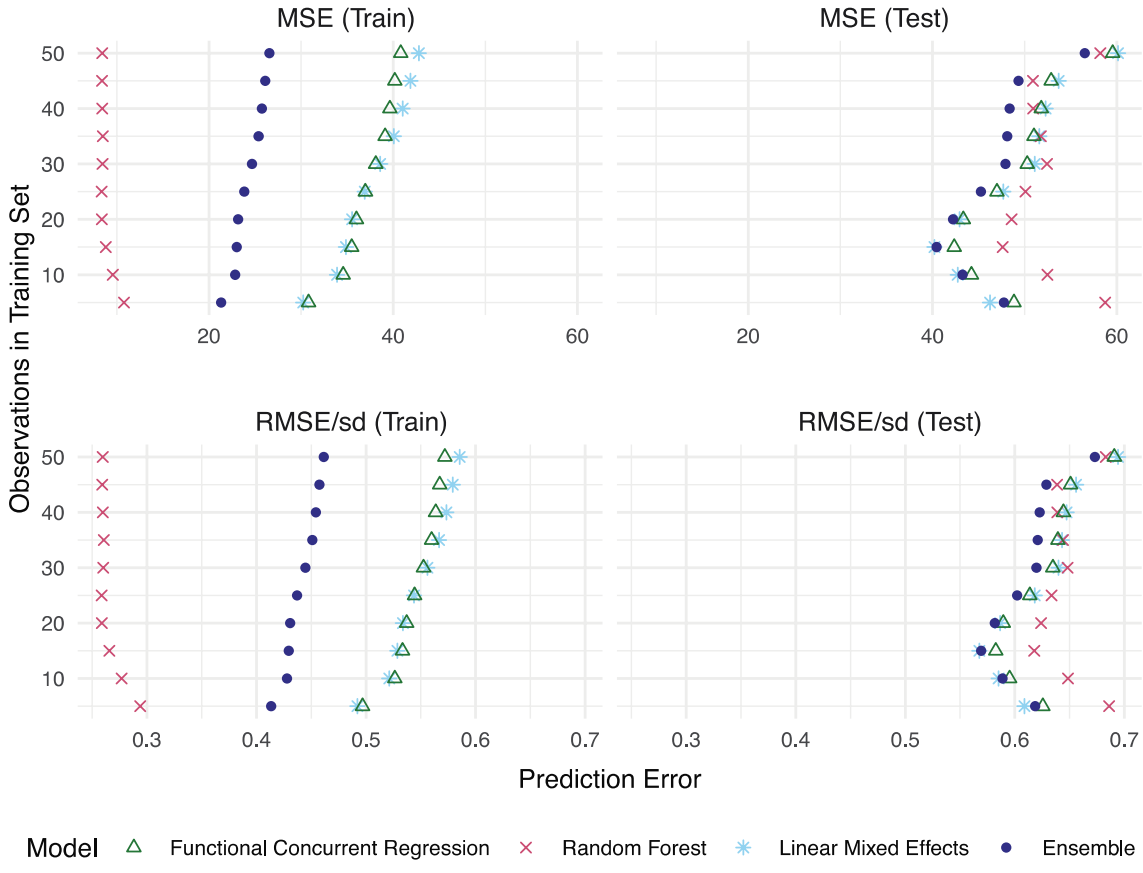


Figure 6. Prediction accuracy among each of the component models and the ensemble. Model performance was measured using the mean squared error (MSE) in predicting LRM-50, described in "Model validation". A standardized measure that can be used to compare prediction accuracy for outcome data with different sizes and magnitudes was obtained by dividing the root MSE (RMSE) by the standard deviation (sd) of the outcome. The RMSE/sd represents the ratio of the model error to the overall variation of the outcome observed in the data.

Discussion

The spaceflight environment is host to a plethora of psychological, operational, and environmental hazards. In this paper, we proposed an ensemble model to predict vigilant attention in astronauts over the course of a space mission. In contrast to previous methods that employed a single prediction method[25] or ensemble[26], ours includes a dynamic component to model time-varying covariate effects. While studies of behavioral health in space or in ground-based space analog environments have traditionally focused on a small number of stressors[7,46,47], our method addresses the intertwined and time-dependent effects of several concurrent stressors. The resulting model flexibly and accurately predicts PVT-B performance. We also identified the most important predictors of behavioral alertness as a combination of individual traits, dynamic psychological state, and environmental conditions.

Ensembles of machine learning models are increasingly popular in human health studies due to their flexibility and accommodation of non-standard data types[48]. Our results suggest that, in settings where the goal is the prediction of a time-varying outcome given a combination of person-level and irregularly measured time series, ensembles which include a functional concurrent regression[32] are able to capture dynamic effects in a powerful way. Furthermore, the incorporation of models with both scalar and functional random effects is useful for individualized predictions (Fig. 8).

Model performance. The ensemble model achieved the best prediction accuracy, outperforming the random forest, LME, and functional regression models in terms of test set MSE. While the random forest had the lowest training MSE, the disparity between its performance on training and test sets suggests that this model may have overfit the data. Interestingly, for most models, the MSE increased as the length of training data t increased, even though we often expect errors to decrease with more training data. One possible explanation is that, for longer training periods, the test set took place later in a person's mission, when the drivers of performance may be different. This may explain why functional regression and LME, which both assume autocorrelation between successive observations, had larger errors in the training set when training time t increased. By contrast, the random forest does not consider the elapsed time in mission.

In space missions, we are often precisely interested in identifying those exceptional times when individuals are at their worst. MSE is a more useful error metric for gauging performance in this case, because it is more sensitive to extreme errors compared to MAE. In our data, higher MSEs were driven by the aforementioned large

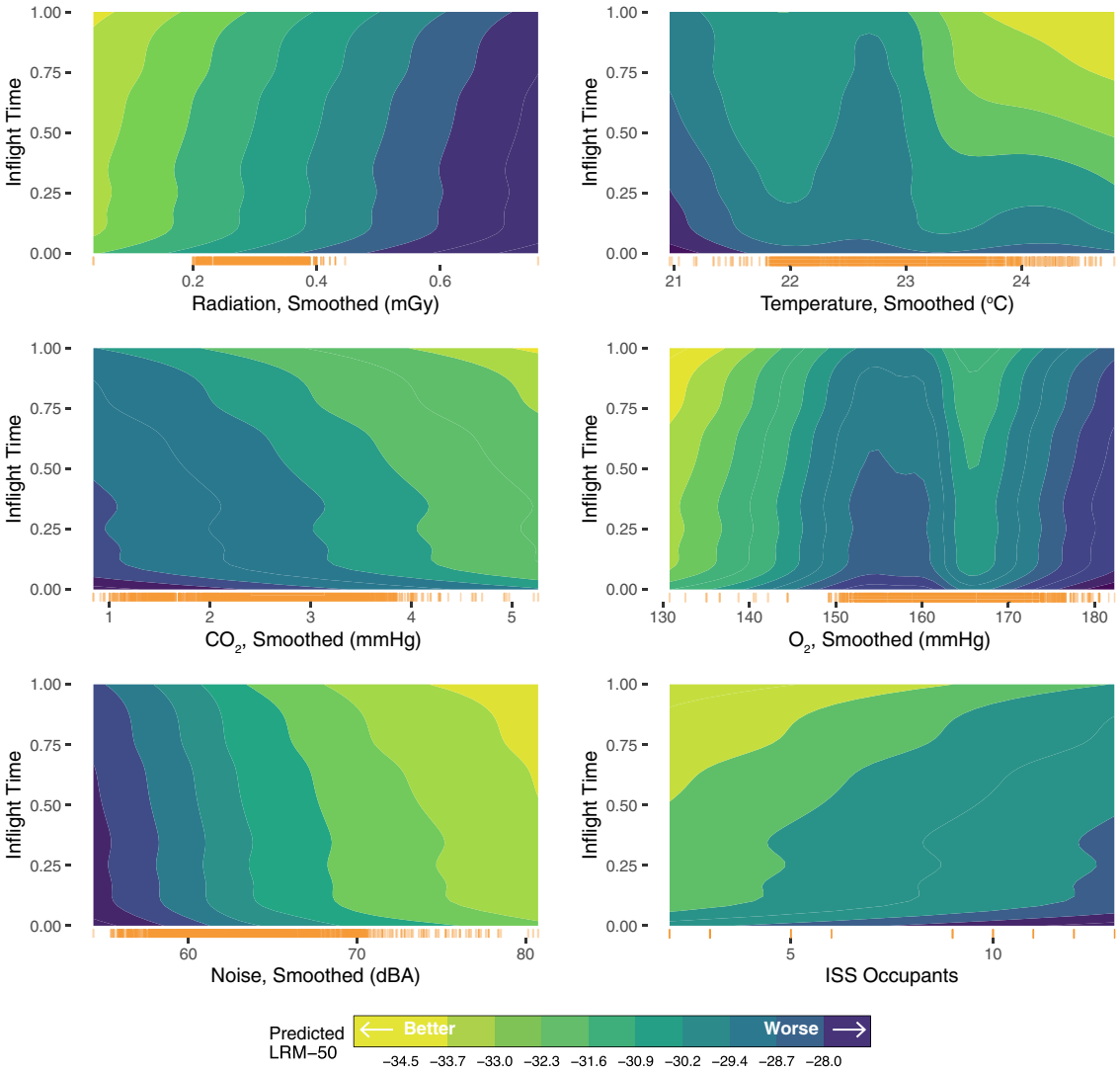


Figure 7. Using predictions from the functional concurrent model, these heat maps show how the non-linear effects of environmental variables on LRM-50 vary over time in mission (ranging from 0 to 1 representing the proportion of mission time elapsed). The same functional concurrent model, which was fit on the entire observed data, was used for each panel. For each environmental variable, predictions were made at a regular grid of time points between 0 and 1, and at all observed values of that environmental variable. All other variables were held at their average (continuous) or reference (categorical) value. The marginal distribution of the environmental variable observations is displayed as a rug plot (orange lines) above the x-axis. We find that better predicted performance (i.e., lower LRM-50, indicated by lighter yellow regions) is generally associated with lower radiation dose, moderate to higher temperatures, higher CO_2, and moderate and lower O_2.

and unanticipated "spikes" in LRM-50, which tended to occur in the latter half of missions. Prediction at these spikes worsened as training length t increased, which may also have contributed to the observed "C" shape when plotting MSE against t (Fig. 6). These observations, together with the superior performance of the ensemble model in terms of both MSE and MAE, imply that the ensemble is a suitable and robust choice for time series in which periods of low variability are interspersed with occasional spikes.

Variable importance. Several predictors ranking highly in importance in our analyses have been previously explored in conjunction with PVT performance, including age[49], sleep duration and time awake[50], ambient temperature[6,51], and lagged PVT performance (performance history)[52,53]. We note that importance is a measure of relative predictive power within the random forest model; EVA events and medication use could have been labeled as "less important" if they did not exhibit sufficient variability in our data, despite their relevance to performance in theory. The distinction between morning and evening RSTs was found to be less important, potentially due to the lack of large differences in LRM-50 distribution by test track (Fig. S10). Any differences were likely due to late evening RST administrations and chronic partial sleep loss in astronauts; this is because vigilant attention is often relatively stable across the first 16 h of the wake period, before deteriorating quickly as a result

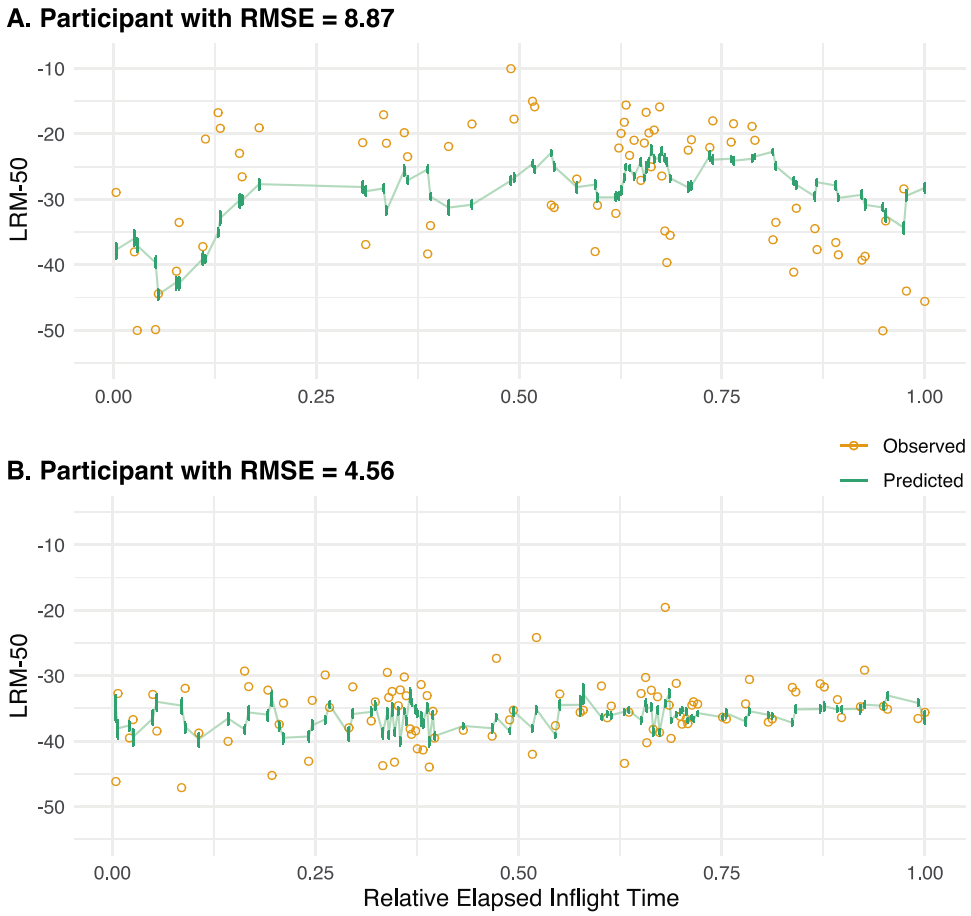


Figure 8. At each time point and for a given astronaut, LRM-50 can be predicted by fitting the model on all preceding data from that astronaut and the full data from other astronauts. The prediction (solid green line) is then made using covariate values from that time point. Sampling variability, given by the height of the error bars at each predicted value, was estimated as the interquartile range of predicted values from 100 bootstrap samples. In order to preserve time-varying trends, individuals were bootstrapped. The actual values of LRM-50 are displayed as hollow yellow circles. To protect astronauts' privacy, the x-axis is the proportion of elapsed mission time rather than calendar time. (A) A participant with prediction error in the highest (worst) 25th percentile (Root mean squared error (RMSE) = 8.87). (B) A participant with prediction error in the lowest (best) 25th percentile (RMSE = 4.56).

of increasing homeostatic pressure and waning circadian promotion of alertness. Finally, collinearity between variables may have affected our results (Fig. S11): for example, poor sleep quality was positively correlated with total sleep missed, which could explain why the former did not rank higher in importance when both variables were in the model. Contrary to prior studies, caffeine was associated with worse predicted performance[54] in the LME analysis, possibly due to its inverse correlation with sleep quality and duration, suggesting that caffeine was consumed because of insufficient sleep but was not able to counteract the neurobehavioral effects of insufficient sleep.

Nonlinear associations of LRM-50 and environmental stressors. Through heatmaps from the functional concurrent regression, we also qualitatively investigated the non-linear effects of environmental conditions on neurobehavioral performance. We found that space radiation is associated with worse performance, which has been established in rodent studies involving a PVT analogue[55,56], but not in humans. The mechanism underlying radiation-related damage to cognitive function is not yet understood[10,12], and this association bears further investigation.

Although the effects of CO_2 exposure on performance are debated[57], we observed a beneficial effect of CO_2 consistent with a recent report in astronaut-like individuals[58]. Temperatures of 23–25 °C were also associated with better predicted LRM-50, consistent with a previous study finding better performance at "cool" temperatures of 26 °C[51]. However, other studies have not detected an effect of temperature on alertness[59]. Also, O_2 exposure levels showed non-linear relationships with PVT performance, where low O_2 concentrations were related to better PVT performance, possibly due to the inverse relationship between O_2 and CO_2. Some of the effects of CO_2, O_2 and temperature may be explained by central nervous system arousing properties of these exposures

once they move outside a range that can easily be accommodated by homeostatic processes (e.g., increase in respiration depth and frequency as well as arousal with high CO_2 concentrations[60]). Finally, as most of the heatmaps indicated, the brighter regions of better LRM-50 performance were predicted to occur at later timepoints when keeping the value of environmental stressor fixed. As individual performance did not generally improve over time, we conjecture that environmental stressors were gradually less coupled with poorest performance, potentially because individuals learned to adjust to the ISS environment. It should be noted that these heatmaps are meant to be hypothesis-generating rather than confirmatory, especially since some of the regions are supported with little data.

Applications, limitations, and future directions. Our findings have three main applications. First, the models were used to identify relevant predictors of objectively assessed alertness via PVT-B in spaceflight. Self-assessments of fatigue and stress, temperature and radiation exposure, caffeine consumption, and past PVT performance were identified as principal correlates of performance. This variable selection can inform space agencies of future areas to concentrate research and mitigation measures.

Second, a tool that can visualize relationships between two predictor variables, such as the R Shiny application (Fig. 4) and functional regression heat maps (Fig. 7), could facilitate the generation of future hypotheses that can later be empirically tested.

Third, exploration-class space missions will involve communication delays and require more crew autonomy. Self-administered tests that assess readiness-to-perform can therefore be a helpful tool in guiding astronaut operational decisions. Using the most current environmental and RST measurements, the predicted LRM-50 score could be incorporated into assessments of astronaut readiness ahead of mission critical tasks and EVAs. For new participants (i.e., individuals whose data did not inform model fitting), the predicted value would be heavily weighted on the group average. This highlights the importance of using a representative sample for model fitting. Our data, which represents one of the largest studies of neurobehavioral performance in astronauts on the ISS, would be a suitable option for making predictions in astronauts, and the R shiny application is a good first step in this direction. However, further validation and tests of astronaut acceptability are required before such a tool could be used in spaceflight.

This study also has several limitations. As the main objective of the ensemble model is to optimize prediction accuracy, the model does not provide statistical inferences on the significance of the effect of any single predictor on the outcome. The coefficients (if they are available) of each component model are not guaranteed to be consistent across models, which may limit interpretability. Second, variable importance rankings are based on permutations or splits of a single variable within a random forest model; therefore, higher-order relationships between multiple predictors and their importance were not assessed. Third, our ensemble prediction weights each model equally, as no model consistently over- or under-performed. Future work could use cross-validation to determine these weights empirically. Fourth, the placement of environmental sensors was constrained by the requirements of spaceflight, and NASA did not collect ambient light data during the study period[61]. We did not consider if the missingness of measurements was itself informative. These factors may have affected the estimated associations between environmental stressors, sleep, and neurobehavioral performance. Fifth, as noted in a recent paper[34], RST observations and sleep–wake measurements were not collected on a daily basis, which limited the ability to measure cumulative sleep loss and assess the effect of circadian misalignments. While our model included many variables previously identified as relevant to PVT performance, there may be other environmental, socioeconomic, or contextual predictors[62] that we are missing. However, it would be straightforward to add new predictors to the existing model. Sixth, the PVT assesses a single cognitive performance domain. While vigilant attention is a prerequisite for many real-world tasks, our findings do not necessarily translate to more complex cognitive or operational tasks. Finally, we did not externally validate model performance on a new group of astronauts, and further work is needed to validate neurobehavioral assessments in spaceflight.

Our statistical methodology has natural extensions. For instance, the ensemble model uses the entire data and concurrent measurements to predict LRM-50. When new observations are made, the entire model must be fit again on the expanded data. An interesting extension could involve Bayesian updating similar to those developed for the unified model of performance[63]. In addition, our model only assesses the direct effects of environmental, psychological, and operational stressors on LRM-50; however, it would also be useful to understand the indirect effects through intermediate variables such as sleep quality and duration. Future work could involve more sophisticated functional mediation analyses that model these relationships explicitly[64]. Finally, stressors such as space radiation[65] and sleep loss[66] may have smaller day-to-day effects on neurobehavioral performance, but significant cumulative effects. While we considered a simple analysis of cumulative radiation, we look forward to study designs and methods that can model the effect of concurrent exposures, as well as exposure duration and history.

Conclusions

To our knowledge, our model is the first investigation of the dynamic, non-linear relationships between common spaceflight stressors, astronaut demographics, and self-reported ratings of sleep and behavioral state on vigilant attention, while also providing individualized predictions of future performance.

The success of spaceflight depends on the physical and mental health of crew members. Our study, based on one of the largest datasets of astronaut neurobehavioral performance and sleep in space, has identified promising avenues in modelling dynamic and personalized profiles of neurobehavioral alertness. Such tools could have important implications for safety and decision-making in one of the most high-profile and dangerous occupations.

Data availability

Per funding agency requirements, the RST data analyzed in this study were uploaded to NASA's Life Sciences Data Archive (LSDA; https://lsda.jsc.nasa.gov) and, together with the environmental data, are available upon request from NASA.

Code availability

All R code required to replicate the analyses herein and produce the R Shiny application are available at https://github.com/danni-tu/TRISH_dynamic_prediction.

Received: 20 March 2022; Accepted: 7 June 2022
Published online: 30 June 2022

References

1. Flynn-Evans, E. E., Barger, L. K., Kubey, A. A., Sullivan, J. P. & Czeisler, C. A. Circadian misalignment affects sleep and medication use before and during spaceflight. *NPJ Microgravity* **2**, 1–6 (2016).
2. Dinges, D. F. An overview of sleepiness and accidents. *J. Sleep Res.* **4**, 4–14 (1995).
3. Åkerstedt, T. Consensus statement: Fatigue and accidents in transport operations. *J. Sleep Res.* **9**, 395 (2000).
4. Mallis, M. M. & DeRoshia, C. W. Circadian rhythms, sleep, and performance in space. *Aviat. Space Environ. Med.* **76**, 94–107 (2005).
5. Jay, S. M., Carley, D. M., Aisbett, B., Ferguson, S. A. & Paterson, J. L. Can stress act as a sleep inertia countermeasure when on-call?. *Biol. Rhythm Res.* **50**, 429–439 (2019).
6. Hudson, A. N., Van Dongen, H. P. A. & Honn, K. A. Sleep deprivation, vigilant attention, and brain function: A review. *Neuropsychopharmacology* **45**, 21–30 (2020).
7. Strangman, G. E., Sipes, W. & Beven, G. Human cognitive performance in spaceflight and analogue environments. *Aviat. Space Environ. Med.* **85**, 1033–1048 (2014).
8. Thirsk, R., Kuipers, A., Mukai, C. & Williams, D. The space-flight environment: The International Space Station and beyond. *Can. Med. Assoc. J.* **180**, 1216–1220 (2009).
9. Stahn, A. C. & Kühn, S. Brains in space: The importance of understanding the impact of long-duration spaceflight on spatial cognition and its neural circuitry. *Cognit. Process.* **22**, 105–114 (2021).
10. Clément, G. R. *et al.* Challenges to the central nervous system during human spaceflight missions to Mars. *J. Neurophysiol.* **123**, 2037–2063 (2020).
11. Tays, G. D. *et al.* The effects of long duration spaceflight on sensorimotor control and cognition. *Front. Neural Circuits* https://doi.org/10.3389/fncir.2021.723504 (2021).
12. Roy-O'Reilly, M., Mulavara, A. & Williams, T. A review of alterations to the brain during spaceflight and the potential relevance to crew in long-duration space exploration. *NPJ Microgravity* **7**, 1–9 (2021).
13. Stahn, A. C. & Kühn, S. Extreme environments for understanding brain and cognition. *Trends Cognit. Sci.* https://doi.org/10.1016/j.tics.2021.10.005 (2021).
14. Basner, M., Mollicone, D. & Dinges, D. F. Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta Astronaut.* **69**, 949–959 (2011).
15. Rupp, T. L., Wesensten, N. J. & Balkin, T. J. Trait-like vulnerability to total and partial sleep loss. *Sleep* **35**, 1163–1172 (2012).
16. Olofsen, E. *et al.* Current approaches and challenges to development of an individualized sleep and performance prediction model. *Sleep (Rochester)* **3**, 24–43 (2010).
17. Borbély, A. A. Two-process model of sleep regulation. *Encycl. Neurosci.* https://doi.org/10.1007/978-3-540-29678-2_6166 (2008).
18. McCauley, P. *et al.* Dynamic circadian modulation in a biomathematical model for the effects of sleep and sleep loss on waking neurobehavioral performance. *Sleep* **36**, 1987–1997 (2013).
19. Postnova, S., Lockley, S. W. & Robinson, P. A. Prediction of cognitive performance and subjective sleepiness using a model of arousal dynamics. *J. Biol. Rhythms* **33**, 203–218 (2018).
20. St. Hilaire, M. A. *et al.* Addition of a non-photic component to a light-based mathematical model of the human circadian pacemaker. *J. Theor. Biol.* **247**, 583–599 (2007).
21. Van Dongen, H. P. A. *et al.* Optimization of biomathematical model predictions for cognitive performance impairment in individuals: Accounting for unknown traits and uncertain states in homeostatic and circadian processes. *Sleep* **30**, 1129–1143 (2007).
22. Graw, P., Kräuchi, K., Knoblauch, V., Wirz-Justice, A. & Cajochen, C. Circadian and wake-dependent modulation of fastest and slowest reaction times during the psychomotor vigilance task. *Physiol. Behav.* **80**, 695–701 (2004).
23. Bhat, S. *et al.* The relationships between improvements in daytime sleepiness, fatigue and depression and psychomotor vigilance task testing with CPAP use in patients with obstructive sleep apnea. *Sleep Med.* **49**, 81–89 (2018).
24. Jewett, M. E., Dijk, D. J., Kronauer, R. E. & Dinges, D. F. Dose-response relationship between sleep duration and human psychomotor vigilance and subjective alertness. *Sleep* **22**, 171–179 (1999).
25. Bermudez, E. B. *et al.* Prediction of vigilant attention and cognitive performance using self-reported alertness, circadian phase, hours since awakening, and accumulated sleep loss. *PLoS ONE* **11**, 1–18 (2016).
26. Cochrane, C., Ba, D., Klerman, E. B. & St. Hilaire, M. A. An ensemble mixed effects model of sleep loss and performance. *J. Theor. Biol.* **509**, 110497 (2021).
27. Blatter, K. & Cajochen, C. Circadian rhythms in cognitive performance: Methodological constraints, protocols, theoretical underpinnings. *Physiol. Behav.* **90**, 196–208 (2007).
28. Williams, D., Kuipers, A., Mukai, C. & Thirsk, R. Acclimation during space flight: Effects on human physiology. *CMAJ* **180**, 1317–1323 (2009).
29. Liu, Q., Zhou, R. L., Zhao, X., Chen, X. P. & Chen, S. G. Acclimation during space flight: Effects on human emotion. *Mil. Med. Res.* **3**, 3–7 (2016).
30. Rencher, A. C. & Schaalje, G. B. Linear models in statistics. *Linear Models Stat.* https://doi.org/10.1002/9780470192610 (2007).
31. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
32. Leroux, A., Xiao, L., Crainiceanu, C. & Checkley, W. Dynamic prediction in functional concurrent regression with an application to child growth. *Stat. Med.* **37**, 1376–1388 (2018).
33. Bergmeir, C. & Benítez, J. M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci. (N.Y.)* **191**, 192–213 (2012).
34. Jones, C. W., Basner, M., Mollicone, D. J., Mott, C. M. & Dinges, D. F. Sleep deficiency in spaceflight is associated with degraded neurobehavioral functions and elevated stress in astronauts on six-month missions aboard the International Space Station. *Sleep* https://doi.org/10.1093/sleep/zsac006 (2022).
35. Basner, M. & Dinges, D. F. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *Sleep* **34**, 581–591 (2011).
36. Barger, L. *et al.* Prevalence of sleep deficiency and hypnotic use among astronauts before, during and after spaceflight: An observational study. *Aviat. Space Environ. Med.* **13**, 904–912 (2014).

37. Basner, M. *et al.* Repeated administration effects on psychomotor vigilance test performance. *Sleep* https://doi.org/10.1093/sleep/zsx187 (2018).
38. Basner, M., Mcguire, S., Goel, N., Rao, H. & Dinges, D. F. A new likelihood ratio metric for the psychomotor vigilance test and its sensitivity to sleep loss. *J. Sleep Res.* **24**, 702–713 (2015).
39. Wotring, V. E. Medication use by U.S. Crewmembers on the International space station. *FASEB J.* **29**, 4417–4423 (2015).
40. Tannenbaum, C., Paquette, A., Hilmer, S., Holroyd-Leduc, J. & Carnahan, R. A Systematic review of amnestic and non-amnestic mild cognitive impairment induced by anticholinergic, antihistamine, GABAergic and opioid drugs. *Drugs Aging* **29**, 639–658 (2012).
41. Marin, R., Cyhan, T. & Miklos, W. Sleep Disturbance in patients with chronic low back pain. *Am. J. Phys. Med. Rehabil.* **85**, 430–435 (2006).
42. Meltzer, E. O. Antihistamine- and decongestant-induced performance decrements. *J. Occup. Environ. Med.* **32**, 327–334 (1990).
43. RCoreTeam. R: A Language and Environment for Statistical Computing (2022).
44. Gómez-Ramírez, J., Ávila-Villanueva, M. & Fernández-Blázquez, M. Á. Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. *Sci. Rep.* **10**, 20630 (2020).
45. Forouzannezhad, P. *et al.* A Gaussian-based model for early detection of mild cognitive impairment using multimodal neuroimaging. *J. Neurosci. Methods* **333**, 108544 (2020).
46. Basner, M. *et al.* Effects of head-down tilt bed rest plus elevated CO_2 on cognitive performance. *J. Appl. Physiol.* **130**, 1235–1246 (2021).
47. Connaboy, C. *et al.* Cognitive performance during prolonged periods in isolated, confined, and extreme environments. *Acta Astronaut.* **177**, 545–551 (2020).
48. Rose, S. Machine learning for prediction in electronic health data. *JAMA Netw. Open* **1**, e181404 (2018).
49. Blatter, K. *et al.* Gender and age differences in psychomotor vigilance performance under differential sleep pressure conditions. *Behav. Brain Res.* **168**, 312–317 (2006).
50. Vetter, C., Juda, M. & Roenneberg, T. The influence of internal time, time awake, and sleep duration on cognitive performance in shiftworkers. *Chronobiol. Int.* **29**, 1127–1138 (2012).
51. te Kulve, M., Schlangen, L. J. M., Schellen, L., Frijns, A. J. H. & van Marken Lichtenbelt, W. D. The impact of morning light intensity and environmental temperature on body temperatures and alertness. *Physiol. Behav.* **175**, 72–81 (2017).
52. Rajaraman, S., Gribok, A. V., Wesensten, N. J., Balkin, T. J. & Reifman, J. Individualized performance prediction of sleep-deprived individuals with the two-process model. *J. Appl. Physiol.* **104**, 459–468 (2008).
53. Olofsen, E., Dinges, D. F. & Van Dongen, H. P. A. Nonlinear mixed-effects modeling: Individualization and prediction. *Aviat. Space Environ. Med.* **75**, A134–A140 (2004).
54. Aidman, E. *et al.* Caffeine may disrupt the impact of real-time drowsiness on cognitive performance: A double-blind, placebo-controlled small-sample study. *Sci. Rep.* **11**, 4027 (2021).
55. Davis, C. M., DeCicco-Skinner, K. L., Roma, P. G. & Hienz, R. D. Individual differences in attentional deficits and dopaminergic protein levels following exposure to proton radiation. *Radiat. Res.* **181**, 258–271 (2014).
56. Cekanaviciute, E., Rosi, S. & Costes, S. Central nervous system responses to simulated galactic cosmic rays. *Int. J. Mol. Sci.* **19**, 3669 (2018).
57. Snow, S. *et al.* Exploring the physiological, neurophysiological and cognitive performance effects of elevated carbon dioxide concentrations indoors. *Build. Environ.* **156**, 243–252 (2019).
58. Scully, R. R. *et al.* Effects of acute exposures to carbon dioxide on decision making and cognition in astronaut-like subjects. *NPJ Microgravity* https://doi.org/10.1038/s41526-019-0071-6 (2019).
59. Mavjee, V. & Home, J. A. Boredom effects on sleepiness/alertness in the early afternoon vs. early evening and interactions with warm ambient temperature. *Br. J. Psychol.* **85**, 317–333 (1994).
60. Langhorst, P., Schulz, B., Schulz, G., Lambertz, M. & Krienke, B. Reticular formation of the lower brainstem. A common system for cardiorespiratory and somatomotor functions: discharge patterns of neighboring neurons influenced by cardiovascular and respiratory afferents. *J. Auton. Nerv. Syst.* **9**, 411–432 (1983).
61. Chellappa, S. L. *et al.* Non-visual effects of light on melatonin, alertness and cognitive performance: Can blue-enriched light keep us alert?. *PLoS ONE* **6**, e16429 (2011).
62. Bessone, P., Rao, G., Schilbach, F., Schofield, H. & Toma, M. The economic consequences of increasing sleep among the urban poor. *Q. J. Econ.* **136**, 1887–1941 (2021).
63. Smith, A. D., Genz, A., Freiberger, D. M., Belenky, G. & Van Dongen, H. P. A. Chapter 8 efficient computation of confidence intervals for bayesian model predictions based on multidimensional parameter space. 213–231. https://doi.org/10.1016/S0076-6879(08)03808-1 (2009).
64. Zeng, S., Rosenbaum, S., Alberts, S. C., Archie, E. A. & Li, F. Causal mediation analysis for sparse and irregular longitudinal data. *Ann. Appl. Stat.* https://doi.org/10.1214/20-AOAS1427 (2021).
65. Maalouf, M., Durante, M. & Foray, N. Biological effects of space radiation on human cells: History, advances and outcomes. *J. Radiat. Res.* **52**, 126–146 (2011).
66. Dinges, D. F. *et al.* Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. *Sleep* **20**, 267–277 (1997).

Acknowledgements

This work is supported by the Translational Research Institute for Space Health through NASA Cooperative Agreement NNX16AO69A (PI: M.B.). The original RST study was supported by the National Aeronautics and Space Administration through NASA NNX08AY09G (PI: DFD). CWJ was supported by a National Institutes of Health NRSA [5T32HL007713]. We thank Betty Lynn Ulrich and the Acoustics Office and Space Radiation Analysis Group at Johnson Space Center for their help with environmental data acquisition. We thank Daniel Mollicone and Chris Mott at Pulsar Informatics Inc. for data collection in the original RST project. We also thank the Lifetime Surveillance of Astronaut Health program and the Life Science Data Archive (LSDA) at NASA for their help with acquiring crew consent and data.

Author contributions

D.T. and H.S. designed the methodology. D.T. implemented code, performed data analyses, and prepared the tables and figures. D.T. and M.B. prepared the main text. M.B., D.F.D., H.S., and C.W.J. contributed to funding acquisition. M.B., D.F.D., C.W.J., E.S.W., A.A.R., and V.E.R. contributed to data acquisition. K.H. and M.K.-L. assisted with project administration. M.B., D.F.D., and H.S. conceptualized and supervised the project. All authors discussed and interpreted the results, and reviewed or contributed to the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-14456-8.

Correspondence and requests for materials should be addressed to H.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dynamic Ensemble Prediction of Cognitive Performance in Spaceflight

Danni Tu [a*], Mathias Basner [b*], Michael G. Smith [b], E. Spencer Williams [c], Valerie E. Ryder [c], Amelia A. Romoser [d], Adrian Ecker [b], Daniel Aeschbach [e,f], Alexander C. Stahn [b], Christopher W. Jones [b], Kia Howard [b], Marc Kaizi-Lutu [b], David F. Dinges [b], Haochang Shou [a+]

[a] Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

[b] Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

[c] Toxicology and Environmental Chemistry, National Aeronautics and Space Administration, Houston, TX, USA

[d] Center for Toxicology and Environmental Health LLC, Houston, TX, USA

[e] Department of Sleep and Human Factors Research, Institute of Aerospace Medicine, German Aerospace Center, Cologne, Germany

[f] Institute of Experimental Epileptology and Cognition Research, Faculty of Medicine, University of Bonn, Bonn, Germany

[*]Equal contribution
[+]Corresponding author at: 219 Blockley Hall, 423 Guardian Drive, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

Email addresses: danni.tu@pennmedicine.upenn.edu (D. Tu), basner@pennmedicine.upenn.edu (M. Basner), hshou@pennmedicine.upenn.edu (H. Shou)

Supporting Tables and Figures

Table S1. All predictors used in the prediction models. (OPS = overall performance score).

Source	*Category*	*Variables*	*Data Type*	*Value*
Reaction Self-Test (RST)	PVT Performance	LRM-50	Time-Varying	Numeric
		Standardized LRM-50	Time-Varying	Numeric
	RST Track Type	Morning/Evening RST	Time-Varying	Binary
		Predicted Lapses	Time-Varying	Numeric
	Sleep	Total Sleep Hours	Time-Varying	Numeric
		Total Sleep Missed (Hours)	Time-Varying	Numeric
	Self-Report	Very Stressed	Time-Varying	Numeric
		Low Workload	Time-Varying	Numeric
		Poor Sleep Quality	Time-Varying	Numeric
		Stress/Fatigue Composite Score	Time-Varying	Numeric
	Medications	Caffeine Doses	Time-Varying	Numeric
		Sleep Aid Flag	Time-Varying	Binary
		Decongestant Flag	Time-Varying	Binary
		Antihistamine Flag	Time-Varying	Binary
		Pain Medication Flag	Time-Varying	Binary
Demographics		Age at Docking	Scalar	Numeric
		Sex	Scalar	Binary
		Average Pre-Flight OPS	Scalar	Numeric
Environmental		Radiation (mGy)	Time-Varying	Numeric
		Temperature (°C)	Time-Varying	Numeric
		Noise (dBA)	Time-Varying	Numeric
		CO_2 (mmHg)	Time-Varying	Numeric
		O_2 (mmHg)	Time-Varying	Numeric
		ISS Occupancy (Count)	Time-Varying	Numeric

Table S2. During the Reaction Self-Test (RST), astronauts rated their behavioral state using 11-point Likert-type rating scales, with prompts and anchors shown below. The first question varied slightly depending on the RST track type (morning or evening).

Post-Sleep (Morning RST) Scales	
1. What was the quality of your sleep?	Anchors: Good ..... Poor
2. How are you feeling right now?	Anchors: Tired ..... Fresh, ready to go
3. How are you feeling right now?	Anchors: Mentally sharp ... Mentally fatigued
4. How are you feeling right now?	Anchors: Energetic ..... Physically exhausted
5. How are you feeling right now?	Anchors: Not stressed ..... Very stressed
6. How are you feeling right now?	Anchors: Not sleepy ..... Very sleepy
Pre-Sleep (Evening RST) Scales	
1. What was today's workload?	Anchors: Very High ..... Very Low
2. How are you feeling right now?	Anchors: Tired ..... Fresh, ready to go
3. How are you feeling right now?	Anchors: Mentally sharp ... Mentally fatigued
4. How are you feeling right now?	Anchors: Energetic ..... Physically exhausted
5. How are you feeling right now?	Anchors: Not stressed ..... Very stressed
6. How are you feeling right now?	Anchors: Not sleepy ..... Very sleepy

Table S3. Principal components analysis (PCA) of the Self-Report 11-point ratings and their loadings on the first principal component. The loadings were used as weights to calculate a composite variable called the stress/fatigue composite score. Based on these weights, higher values of the composite score correspond to greater feelings of stress, physical and mental tiredness and fatigue, and feelings of sleepiness. Due to its low correlation with other variables, the workload was not highly weighted in this score.

Variable	***Loading***
Workload (0 = high, 10 = low)	-0.0449
Sleep Quality (0 = good, 10 = poor)	0.2295
Feeling Sleepy (0 = not at all, 10 = very much)	0.4618
Physically Exhausted (0 = energetic, 10 = physically exhausted)	0.4898
Mentally Fatigued (0 = mentally sharp, 10 = mentally fatigued)	0.4607
Tiredness (0 = tired, 10 = fresh, ready to go)	-0.4812
Stress (0 = not stressed, 10 = very stressed)	0.2198

Table S4. To match RST observations to corresponding values of environmental and operational variables, we first summarized variables in terms of their daily value (if the variables were measured daily) or hourly average (if measured more frequently). When temperature and noise data were collected at multiple sensors simultaneously, efforts were made to "location match" the data when possible, by using the sensor measurement in the same module where the RST was taken (see Section 2.6). LOESS interpolation was used for radiation, temperature, CO_2, and O_2. Linear interpolation was used for noise.

Variable	*Frequency of Collected Measurements*	*Value Used with RST Observation*	*Location Matching (if available)*	*Interpolated Value*
Radiation (mGy)	Daily	Daily value	No	LOESS
Temperature (°C)	Multiple times per minute	Hourly average	Yes	LOESS
Noise (dBA)	Once per minute	Hourly energetic average	Yes	Linear
CO_2 (mmHg)	Multiple times per minute	Hourly average	No	LOESS
O_2 (mmHg)	Multiple times per minute	Hourly average	No	LOESS
ISS Occupancy (Count)	Daily	Daily value	N/A	N/A

Table S5. Rates of missingness for time-varying variables. The percentage of missing observations was calculated as the number of in-flight RST observations taken at times where the corresponding value of the environmental variable (i.e., the daily average of radiation or ISS occupants, and the hourly average of noise, temperature, CO_2, and O_2) was available, divided by the total number of in-flight RST observations (n = 2094). Notably, almost all values of noise were missing, since that variable was only recorded on 47 unique 24-hour measurement periods throughout the study period. Sleep diary data (i.e., bedtime, time taken to fall asleep, and duration of asleep) was only collected during the morning RST, so the proportion of missing observations was calculated out of the 1,105 in-flight morning RSTs.

Variable	**Missing Observations (%)**
Radiation Dose (mGy)	5.40
Noise (dBA)	99.71
Temperature (°C)	3.63
CO_2 (mmHg)	7.02
O_2 (mmHg)	22.97
ISS Occupants (Count)	0.00
Sleep Diary	10.68

Table S6. In testing data, the ensemble model outperforms all component models in terms of averaged mean squared error (MSE) (see Section 2.8 for details). Values in parentheses represent the interquartile range of 25th and 75th percentiles. The model trained on the full set of covariates ("All") performed best, but performance was similar when retaining only the Top 10 variables (Section 3.1). Including the O_2, CO_2, and noise variables did not drastically alter performance, although models excluding noise performed the worst overall.

Covariates	Including Noise Variable	Variable	Linear Mixed Effects	Random Forest	Functional Concurrent Regression	Ensemble
All	Yes	MSE (Test)	48.86 (43.77, 52.11)	52.18 (50.3, 52.47)	49.13 (44.92, 51.61)	**46.92 (43.76, 48.31)**
Top 10	Yes	MSE (Test)	49.34 (45.14, 52.51)	51.48 (49.71, 51.63)	50.28 (47.57, 52.1)	**47.60 (45.24, 49.16)**
Top 10 + CO_2 + O_2	Yes	MSE (Test)	49.48 (45.18, 52.52)	51.76 (50.1, 52.1)	50.20 (47.3, 52.02)	**47.52 (45.35, 49.14)**
Top 10	No	MSE (Test)	49.82 (46.12, 52.29)	51.06 (49.28, 50.98)	50.97 (48.6, 52.82)	**47.98 (45.67, 49.18)**
Top 10 + CO_2 + O_2	No	MSE (Test)	49.81 (46.02, 52.16)	51.49 (49.18, 51.75)	50.74 (48.2, 52.49)	**48.04 (45.83, 49.26)**
All	Yes	MSE (Train)	37.57 (35.03, 40.78)	**8.80 (8.39, 8.72)**	37.17 (35.62, 39.51)	24.26 (23.05, 25.66)
Top 10	Yes	MSE (Train)	38.74 (36.31, 41.85)	**9.87 (9.57, 9.75)**	39.87 (38.58, 42.1)	26.00 (24.86, 27.38)
Top 10 + CO_2 + O_2	Yes	MSE (Train)	38.73 (36.3, 41.82)	**9.09 (8.68, 9.04)**	39.57 (38.28, 41.8)	25.47 (24.42, 26.75)
Top 10	No	MSE (Train)	39.54 (37.27, 42.58)	**9.69 (9.4, 9.53)**	40.65 (39.39, 42.84)	26.36 (25.26, 27.7)
Top 10 + CO_2 + O_2	No	MSE (Train)	39.47 (37.2, 42.49)	**9.62 (9.31, 9.48)**	40.37 (39.1, 42.59)	26.25 (25.16, 27.55)

Table S7. Assuming that covariate values are known at future timepoints, model performance remains acceptable when predicting more than 1 timepoint ahead. Prediction performance was assessed using the same method as described in Section 2.8, but we allowed the size of the test set to range from 1 to 7 future observations. Values in parentheses represent the interquartile range of 25th and 75th percentiles. Because the RST was administered twice a day approximately every 4 days, these future predictions corresponded to predictions of around 2 to 15 days ahead on average. Cell values are the mean squared error in the testing data averaged over participants, then length of training days t. Overall, we find that the errors increase slowly as the prediction horizon grows.

Future Observations in Test Set	Average Length of Prediction Horizon (Days)	Linear Mixed Effects	Random Forest	Functional Concurrent Regression	Ensemble
1	1.88	48.86 (43.77, 52.11)	52.18 (50.3, 52.47)	49.13 (44.92, 51.61)	**46.92 (43.76, 48.31)**
2	4.75	49.86 (44.84, 52.77)	53.48 (51.76, 53.85)	50.03 (45.69, 52.13)	**47.92 (44.53, 49.05)**
3	6.55	50.17 (45.86, 52.91)	53.80 (51.31, 54.19)	49.95 (46.12, 51.98)	**48.03 (45.11, 49.06)**
5	11.34	50.86 (48.01, 53.18)	53.87 (51.8, 54.67)	49.78 (47.55, 51.73)	**48.05 (46.5, 49.47)**
7	15.32	51.69 (49.23, 53.88)	54.49 (53.02, 54.99)	50.31 (48.14, 52.24)	**48.57 (47.22, 49.56)**

Table S8. Coefficients for the linear mixed effects model to predict LRM-50 score, with a random intercept for each participant and AR1 correlation structure. To enable comparisons of coefficients between variables, both the numeric covariates and the outcome were *z*-scored (i.e., linearly scaled to have a mean of 0 and a standard deviation of 1). Positive coefficients are associated with an increase in LRM-50 (worse performance); negative coefficients are associated with a decrease in LRM-50 (better performance). Due to low numbers of EVAs, neither EVA flag was included in the model. A limitation of this model is that it can only accommodate linear relationships between predictors and the outcome, which may not reflect their actual relationship. Note: *$p < 0.05$; **$p < 0.01$.

Variable	**Coefficient (95% CI)**
(Intercept)	-0.305 (-1.011, 0.400)
Radiation, Smoothed (mGy)	0.011 (-0.029, 0.052)
Noise, Smoothed (dBA)	-0.090 (-0.120, -0.059)**
CO_2, Smoothed (mmHg)	-0.014 (-0.052, 0.023)
O_2, Smoothed (mmHg)	-0.001 (-0.036, 0.034)
ISS Occupants	0.018 (-0.016, 0.053)
Temperature, Smoothed (°C)	-0.061 (-0.111, -0.011)*
Sex = Male	0.305 (-0.555, 1.164)
Age at Dock	-0.082 (-0.428, 0.264)
Average Pre-flight OPS	-0.207 (-0.514, 0.100)
Sleep Aid Flag	0.058 (-0.077, 0.192)
Antihistamine Flag	0.010 (-0.017, 0.037)
Morning RST	0.107 (0.046, 0.168)**
Stress/Fatigue Composite Score	0.179 (0.143, 0.215)**
Low Workload	-0.019 (-0.050, 0.012)
Poor Sleep Quality	-0.005 (-0.046, 0.035)
Very Stressed	-0.017 (-0.065, 0.030)
Total Sleep Hours	-0.059 (-0.096, -0.023)**
Total Sleep Missed	-0.001 (-0.037, 0.034)
Predicted Lapses	0.003 (-0.038, 0.045)
Caffeine Doses	0.020 (-0.035, 0.074)
Decongestant Flag	0.104 (-0.135, 0.344)
Pain Medication Flag	-0.088 (-0.229, 0.053)
LRM-50 (Lagged)	0.121 (0.084, 0.157)**

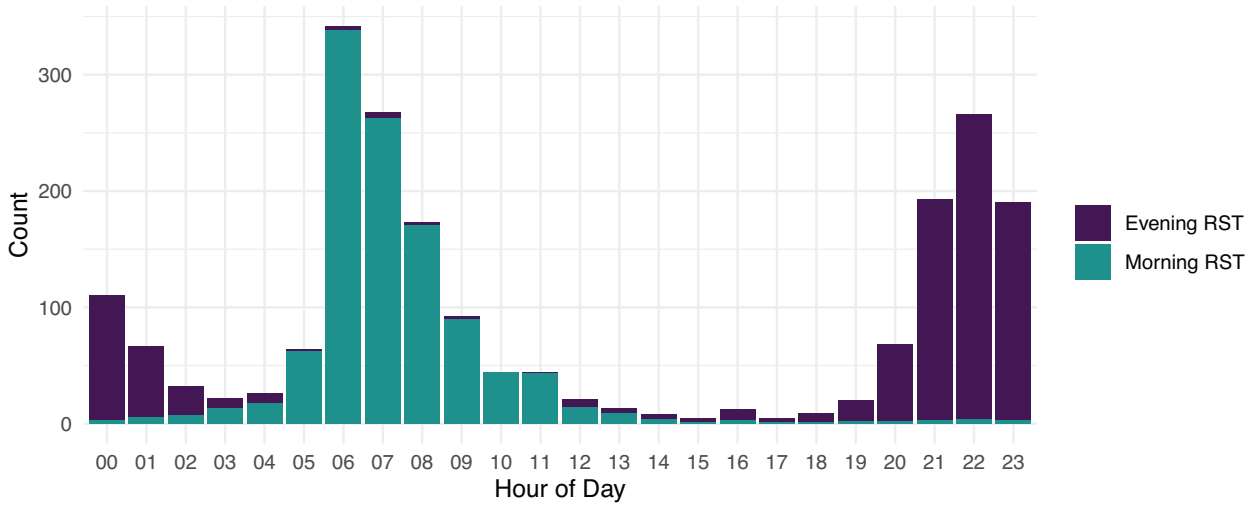
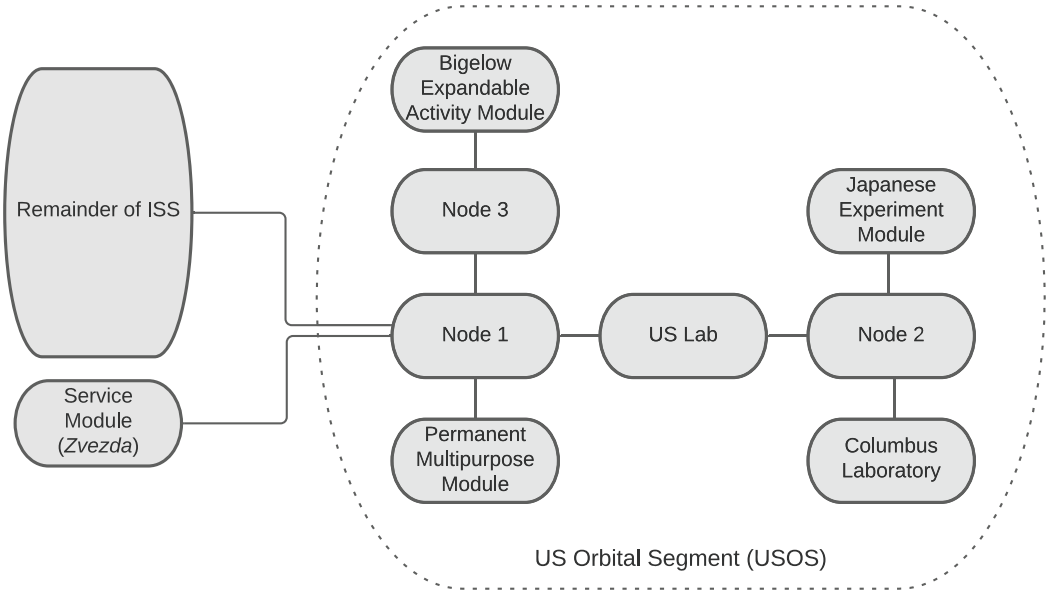


Figure S1. In-flight RST observations occurred at all hours of the day, though morning RSTs (which were taken after awakening) were concentrated in the morning hours and evening RSTs (taken before bedtime) in the nighttime hours.

A. Schematic of ISS Modules



B. Locations of MCA Sample Ports Aboard ISS

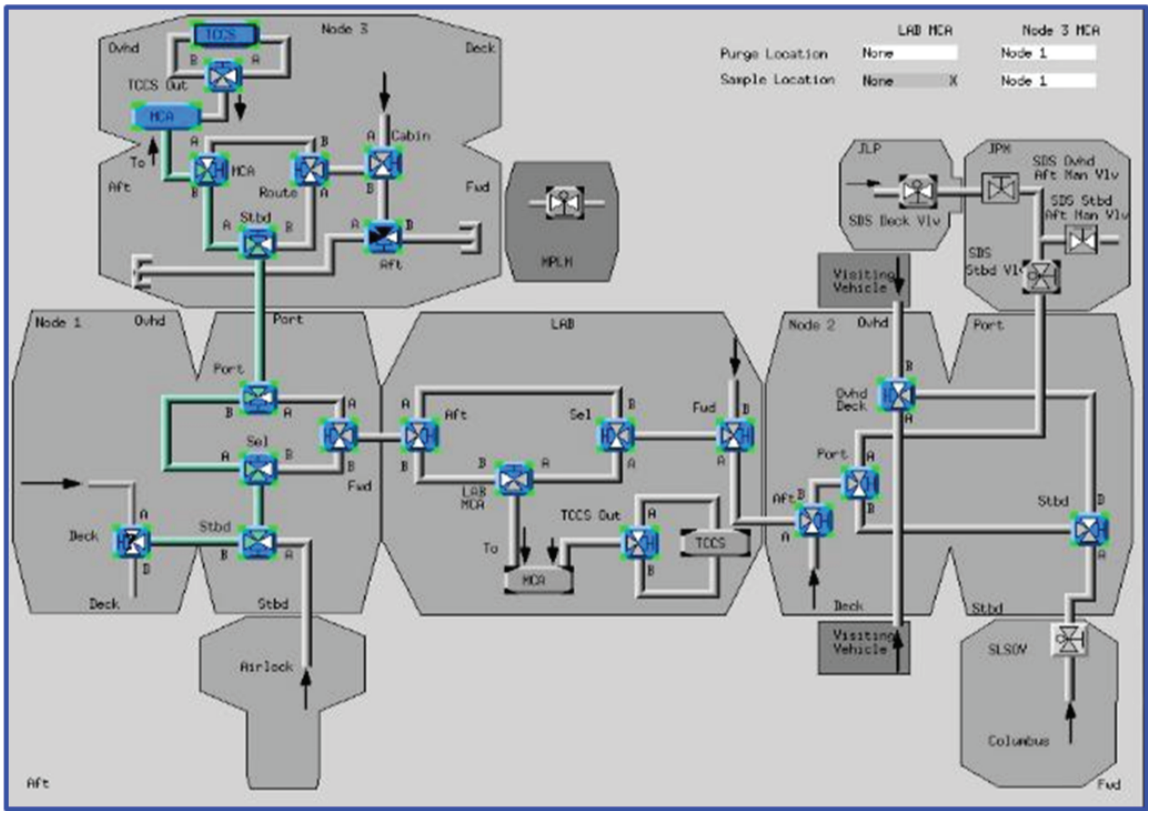


Figure S2. Modules of the International Space Station (ISS). Panel A: schematic of the ISS modules in the US Orbital Segment (USOS), where the environmental data used in this study was recorded. Most of the in-flight RST observations occurred in the US and Node 2 modules (17.4% and 75.5%, respectively). Panel B: locations of the Major Constituent Analyzer (MCA) sample ports (shown as blue squares with a blue cross and 4 green dots) throughout the USOS modules. These ports sampled CO_2 and O_2 (and other atmospheric components) into mass spectrometer-based MCA units located in Node 3 and US Lab.

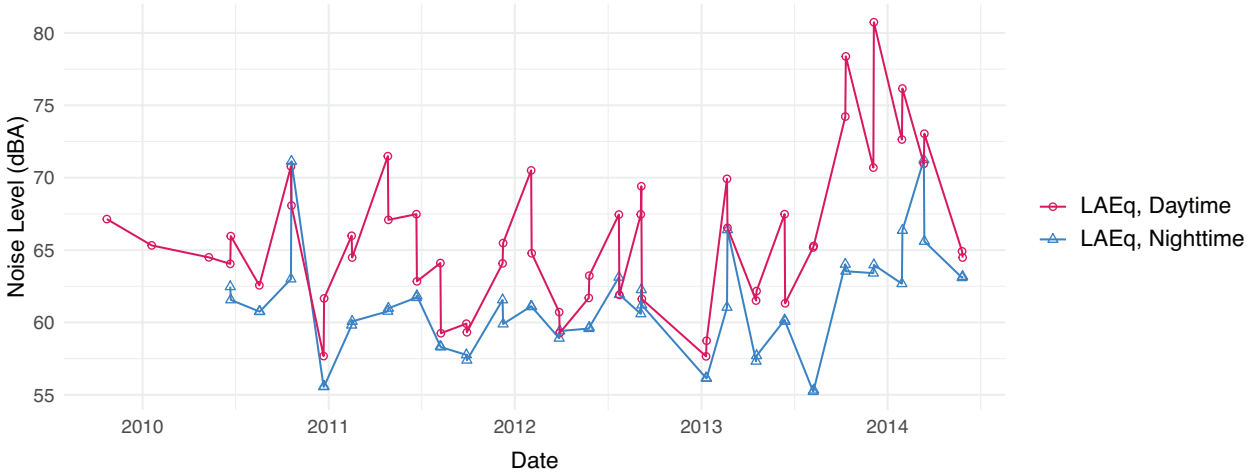


Figure S3. Average noise was higher during daytime hours (7:00 to 22:59) than nighttime hours (23:00 to 6:59). In the study period, noise levels in A-weighted decibels (dBA) were measured during 47 occasions lasting approximately 24 hours each. Red circles correspond to daytime noise measurements, averaged using the energetic average ($L_{A,eq}$). Blue triangles correspond to nighttime measurements.

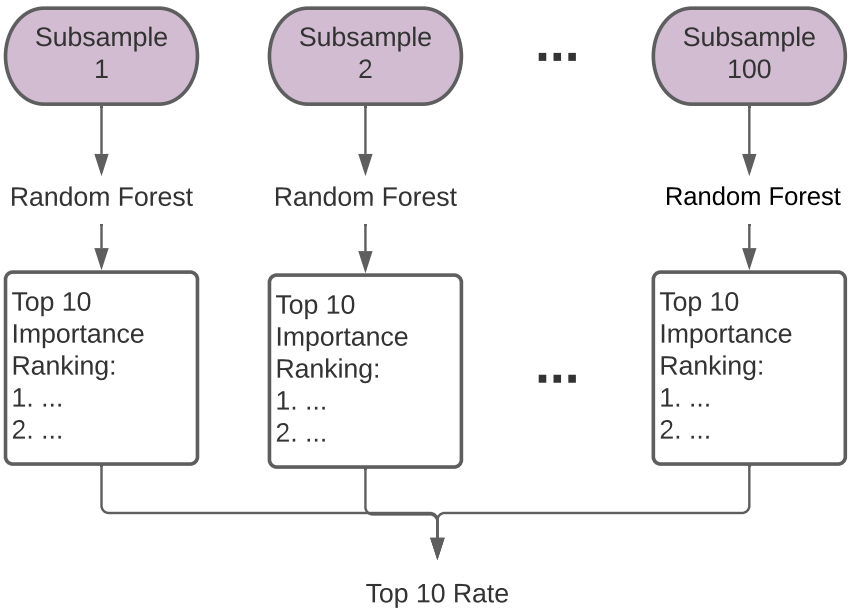


Figure S4. Variable importance defined in Section 2.9 was calculated by repeatedly fitting random forest models to resampled subsets of the data. For each subsample, the corresponding importance ranking given by the %IncMSE metric was determined. Then, the overall Top 10 Rate for each variable was calculated as the proportion of subsamples where that variable was among the 10 most important.

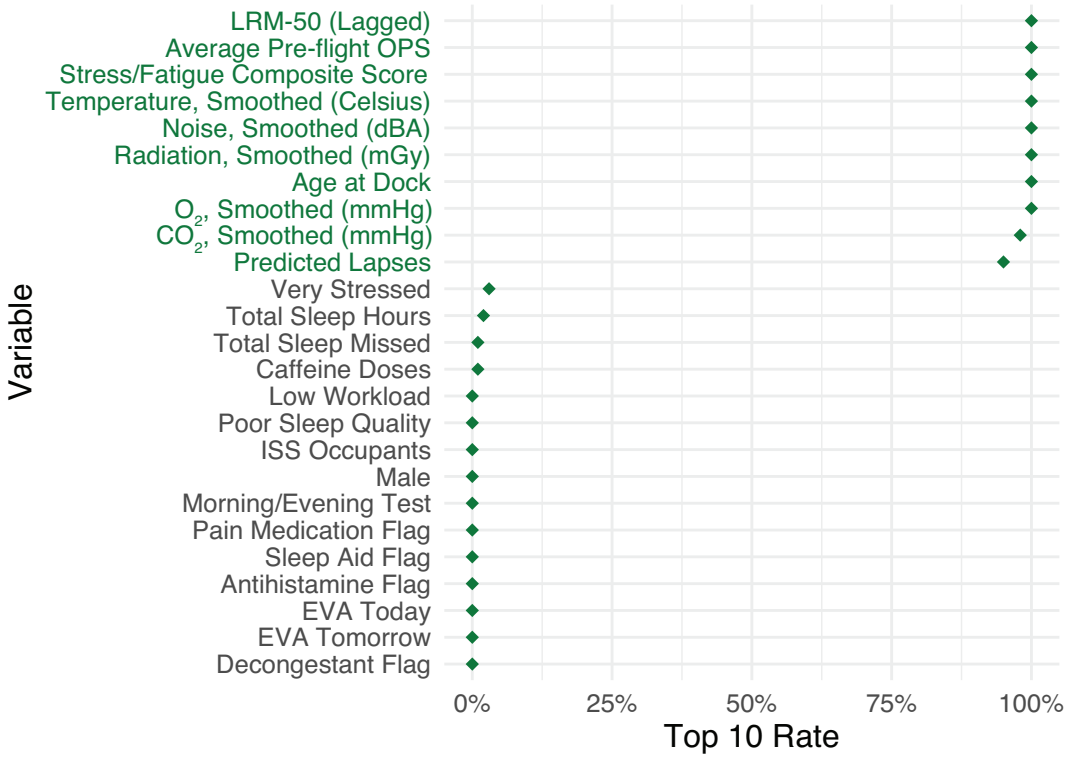


Figure S5. We also considered assessing variable importance by the decrease in node impurity, as measured by the Gini coefficient after splitting on that variable, averaged over all trees in a random forest. To understand which variables were consistently important in different subsamples of the data, variable importance rankings were obtained from 100 resampling draws. The resulting "Top 10 Rate" (x-axis) describes how a given variable, over resampling trials, is repeatedly among the 10 most important variables in a random forest model. We then defined the Top 10 variables as those which scored higher on this metric; these consisted of the lagged LRM-50 through the number of predicted lapses (green text). Compared to importance measured by the increase in MSE (Figure 5), node impurity tends to have less variability between resampling iterations. The top 7 variables in terms of node purity (lagged LRM-50 through age) were the same as those from the increase in MSE. (OPS = Overall Performance Score; RST = Reaction Self-Test; ISS = International Space Station; EVA = extravehicular activity.)

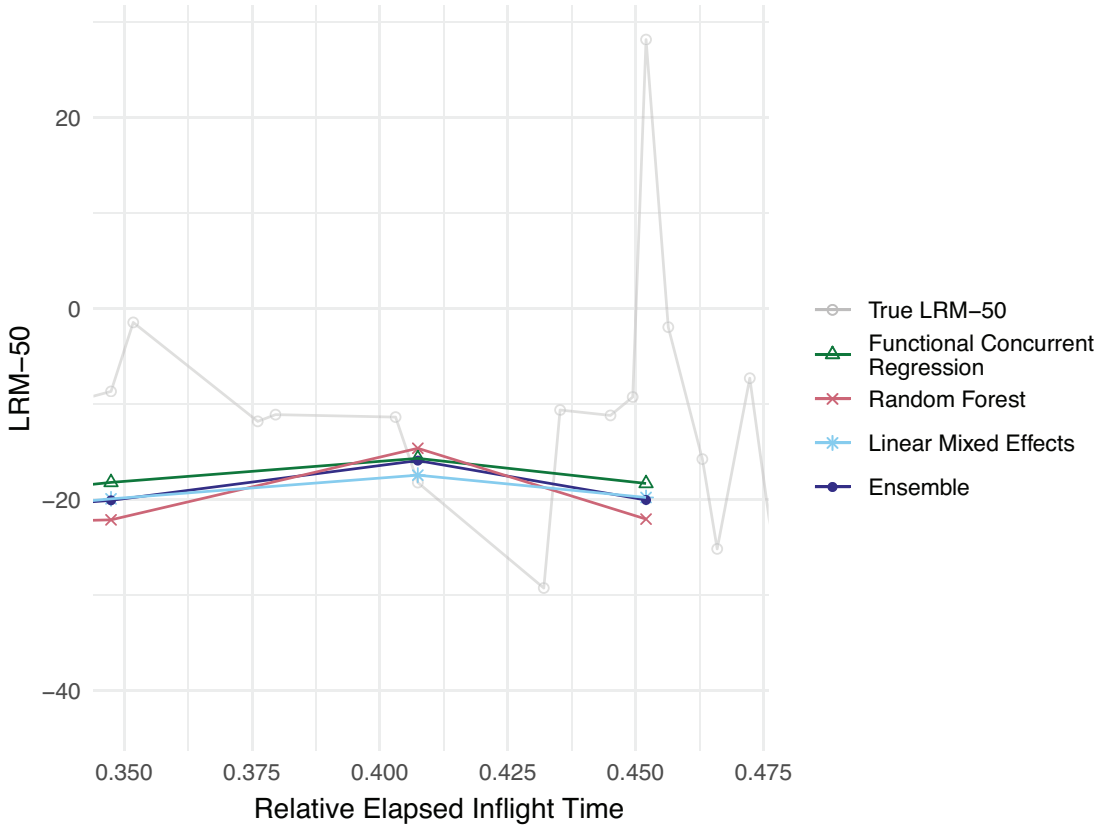


Figure S6. Example of an extreme "spike" in observed LRM-50 from one participant. The light gray hollow circles and line correspond to actual LRM-50 observations. The points and lines in color represent the model predictions at the spike and 5 or 10 observations previous. To protect astronauts' privacy, the x-axis is the proportion of elapsed mission time rather than calendar time. At the spike, which occurs just after time = 0.45, the true LRM-50 = 28.14, while more recent values ranged from -30 to -5. The models perform adequately at previous points, but are not able to anticipate the spike. This disparity between predicted values and the true value resulted in large MSEs of 2000-2500.

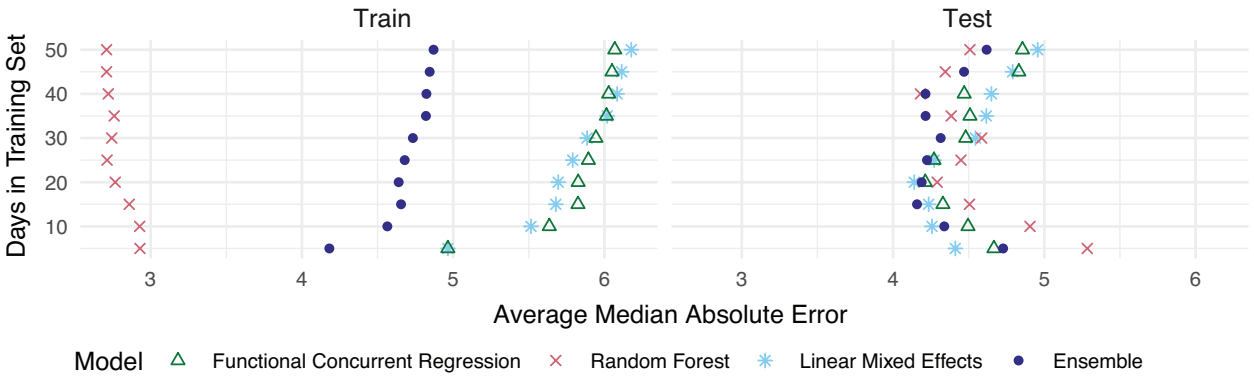


Figure S7. Because the mean squared error (MSE) tends to heavily penalize large errors in predicting LRM-50, we also compared the models in terms of the *median* absolute error, which is more robust to extreme values (i.e., by penalizing them less). The median absolute error (MAE) was calculated using a similar aggregation procedure as in Section 2.8, but replacing the $MSE(i,t)$ with $MAE(i,t)$ defined in Section 3.2. The scale of median absolute errors was smaller than that of square-root MSEs, implying that the latter were likely influenced by larger errors. The ensemble model continued to out-perform the other models on average, though its lead over other models is narrower in terms of the MAE.

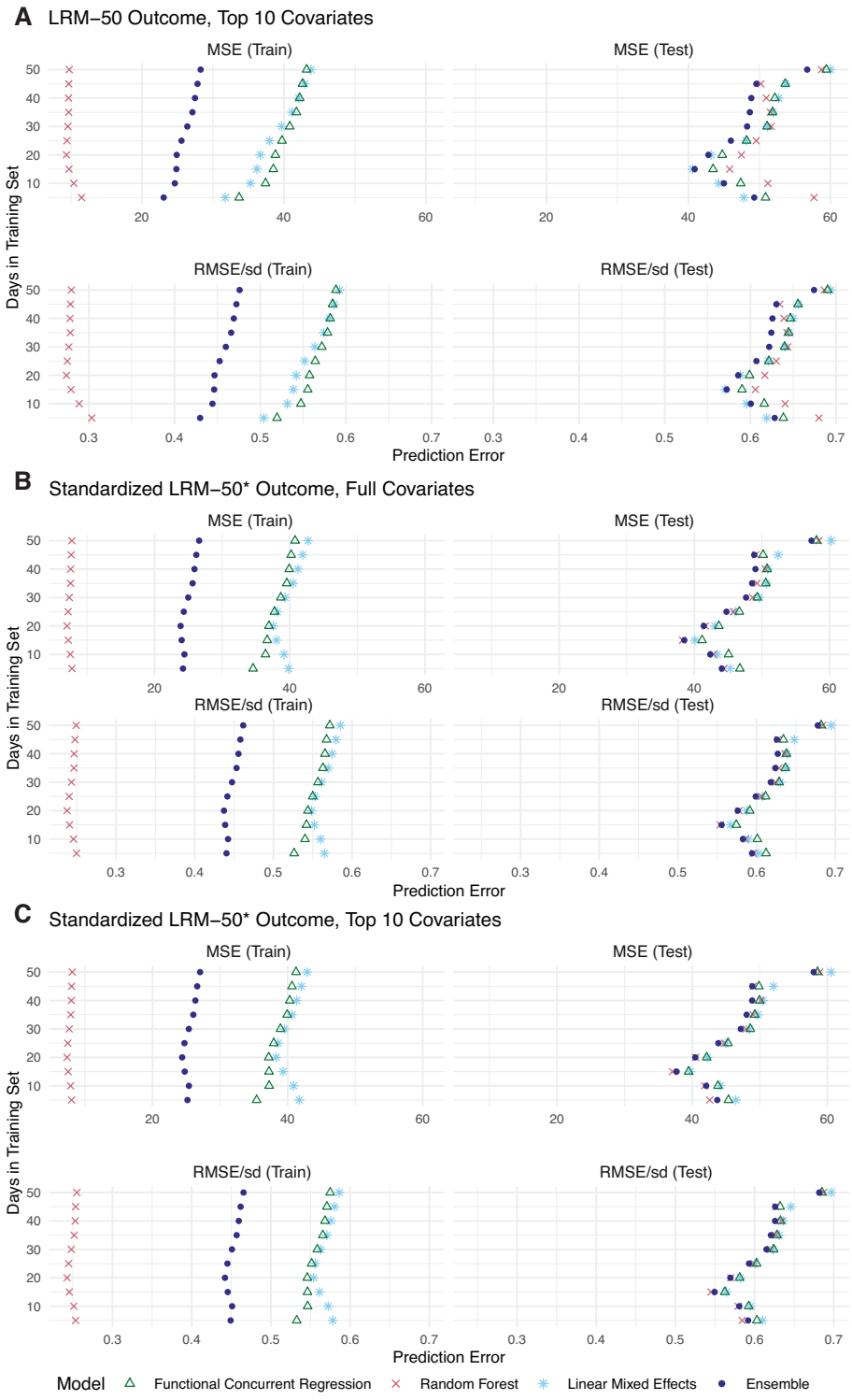


Figure S8. The ensemble model again outperformed the 3 component models under various settings. Panel A: only the top importance variables (defined by the union of the environmental variables and the 10 variables most frequently ranked in the Top 10 in variable importance; see Figure 5) were included as predictors of LRM-50. Panels B and C: we also considered the standardized LRM-50 outcome, calculated as the LRM-50 score linearly scaled by each participant's mean and standard deviation of LRM-50. *For the standardized LRM-50 outcome, the prediction errors were transformed back to the original scale to enable comparisons.

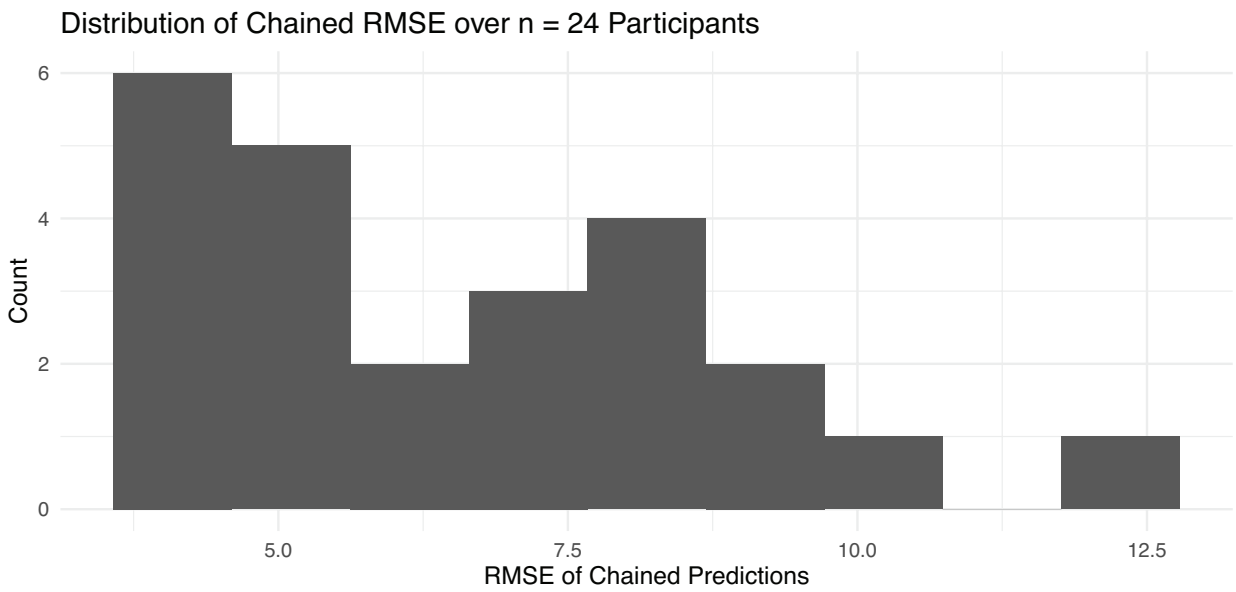


Figure S9. At any point in time, the most up-to-date information can be used to predict the next LRM-50 score, assuming that all covariate values at that future point are known. Using the procedure outlined in Section 3.4, we obtain a series of "chained" predictions for each participant. This histogram shows the distribution of the average chained RMSE among the 24 participants in our study. The mean RMSE = 6.56 (sd = 2.19) and the median RMSE = 6.12 (IQR = 4.78 to 8.02).

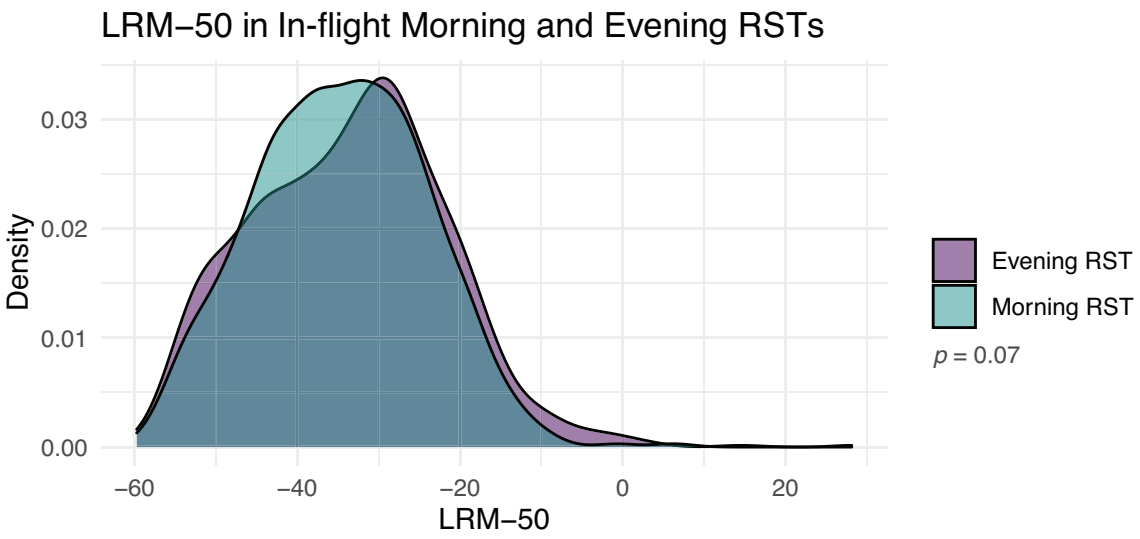


Figure S10. The distribution of LRM-50 scores in morning (n = 1105) and evening (n = 989) in-flight RST observations. The Welch two-sample *t*-test for difference in means between LRM-50 values from morning and evening tests had a *p*-value of 0.07 (*t* statistic = -1.84, *df* = 2006).

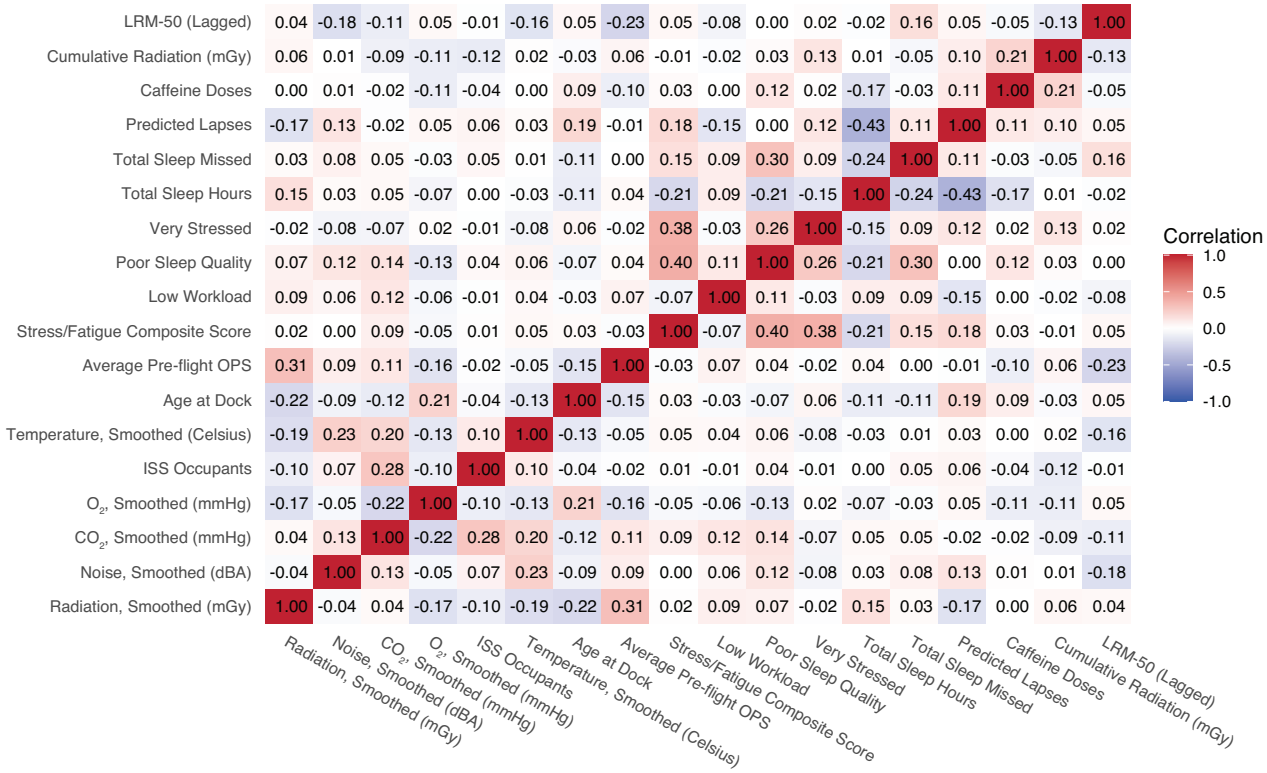


Figure S11. Heatmap of Pearson correlation between all numeric variables included in the ensemble model.

Supplementary Methods 1: Cumulative Radiation Dose

Cumulative radiation dose was calculated for each person on each in-flight day of their mission, using the cumulative sum of (smoothed) daily radiation dose in mGy.

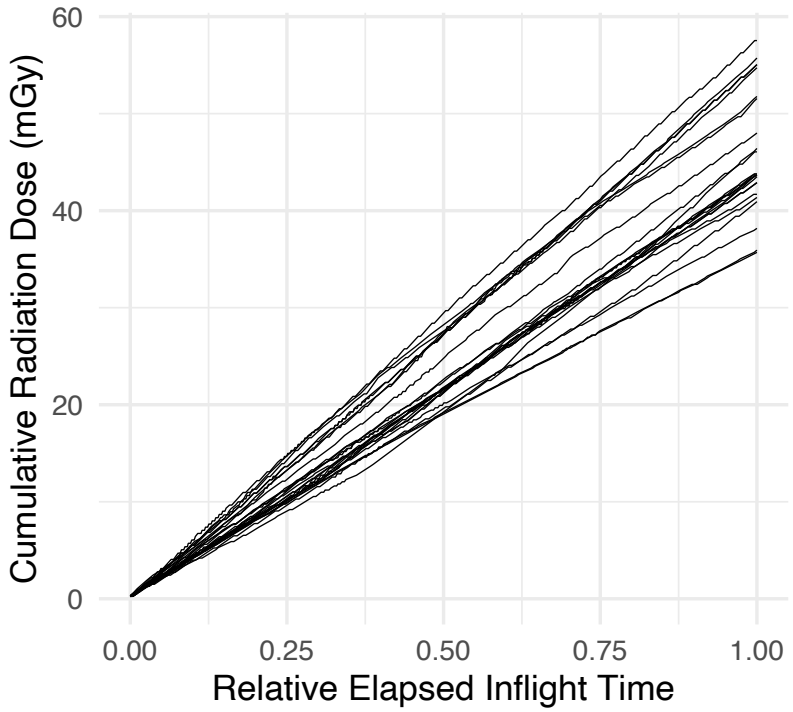


Figure S12. Cumulative radiation dose trajectories for all 24 astronauts in our data.

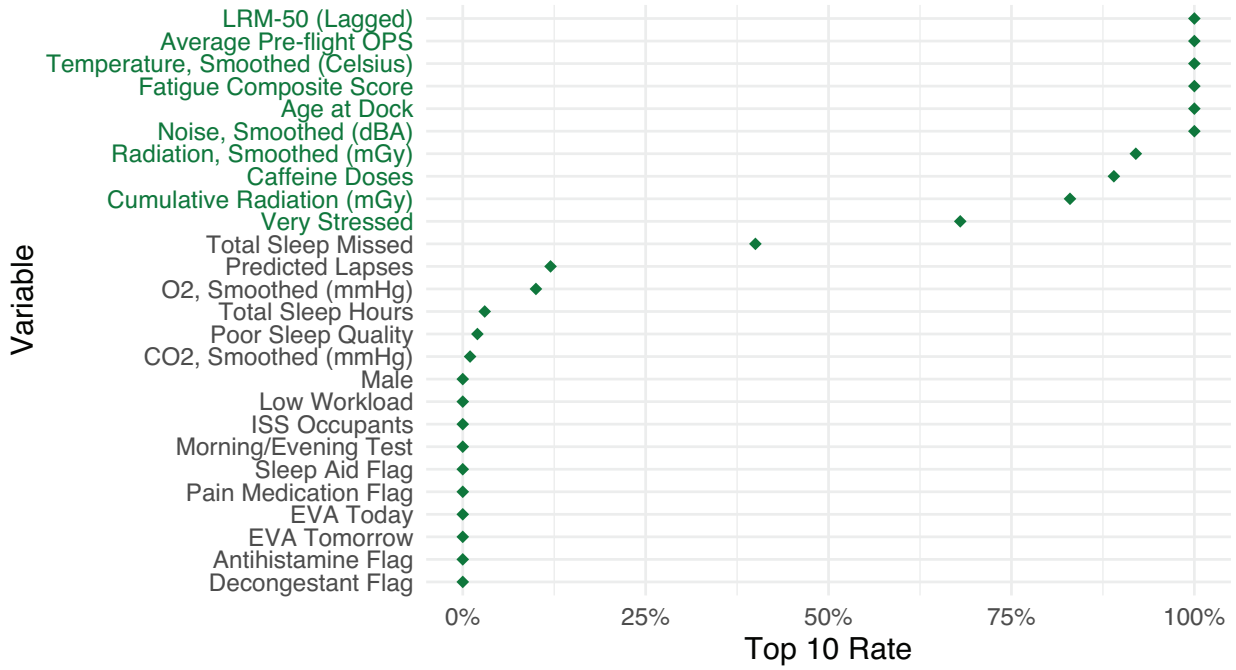


Figure S13. Random forest variable importance defined by the increase in MSE (Section 2.9) for a model including cumulative radiation (mGy). Both the concurrent radiation dose and cumulative radiation were found to be important predictors of LRM-50.

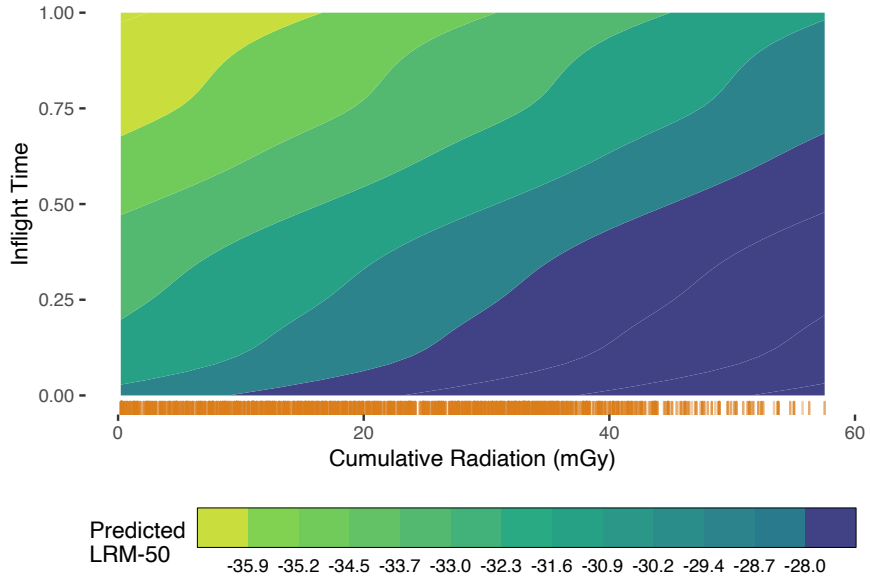


Figure S14. Using functional concurrent regression models containing cumulative radiation dose (mGy), we again produced heatmaps of the association between inflight time, cumulative radiation dose, and predicted LRM-50 (Section 3.3). The marginal distribution of the environmental variable observations is displayed as a rug plot (orange lines) above the x-axis. Similar to our findings for daily radiation dose (mGy), less cumulative radiation is associated with better predicted performance.

Table S9. Including cumulative radiation as a covariate did not dramatically alter model performance in terms of averaged mean squared error (MSE) (see Section 2.8 for details). Values in parentheses represent the interquartile range of 25th and 75th percentiles. The "Top 10" variables were defined by the 10 most important variables (Figure S13), which included cumulative radiation. As before, the model trained on the full set of covariates ("All") performed best, but performance was similar when retaining only the new set of Top 10 variables.

Covariates	Including Noise Variable	Variable	Linear Mixed Effects	Random Forest	Functional Concurrent Regression	Ensemble
All	Yes	MSE (Test)	48.82 (44.03, 51.83)	52.16 (50.38, 52.67)	49.37 (45.07, 52.1)	**46.98 (43.95, 48.56)**
Top 10	Yes	MSE (Test)	49.26 (45.09, 52.22)	51.38 (49.89, 52.2)	50.41 (47.57, 52.61)	**47.44 (45.28, 48.96)**
Top 10 + CO_2 + O_2	Yes	MSE (Test)	49.37 (45.13, 52.15)	51.85 (50.11, 52.27)	50.37 (47.38, 52.48)	**47.43 (45.25, 48.95)**
Top 10	No	MSE (Test)	49.73 (45.99, 52.11)	51.36 (49.54, 51.3)	50.48 (47.83, 52.1)	**47.74 (45.52, 48.82)**
Top 10 + CO_2 + O_2	No	MSE (Test)	49.59 (45.96, 52.03)	51.11 (49.28, 52.65)	50.66 (48.05, 52.55)	**47.71 (45.36, 49.16)**
All	Yes	MSE (Train)	37.64 (35.05, 40.93)	**8.71 (8.33, 8.61)**	36.72 (34.95, 39.26)	24.15 (22.88, 25.61)
Top 10	Yes	MSE (Train)	38.80 (36.31, 42)	**9.50 (9.17, 9.38)**	39.63 (38.22, 42.06)	25.71 (24.57, 27.11)
Top 10 + CO_2 + O_2	Yes	MSE (Train)	38.80 (36.31, 41.98)	**8.93 (8.54, 8.86)**	39.41 (37.98, 41.83)	25.31 (24.21, 26.64)
Top 10	No	MSE (Train)	39.54 (37.2, 42.66)	**9.35 (9.01, 9.18)**	40.12 (38.71, 42.51)	25.99 (24.86, 27.35)
Top 10 + CO_2 + O_2	No	MSE (Train)	39.54 (37.2, 42.66)	**9.60 (9.32, 9.48)**	40.41 (38.97, 42.83)	26.18 (25.01, 27.62)