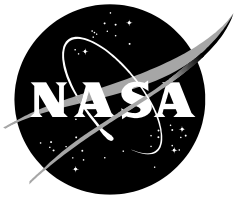
NASA/TP-20210022966



Dynamic Ensemble Prediction of Cognitive Performance in Space

Danni Tu
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Mathias Basner
Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Michael G. Smith
Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

E. Spencer Williams
Johnson Space Center, Houston TX

Valerie E. Ryder
Johnson Space Center, Houston TX

Amelia A. Romoser
Center for Toxicology and Environmental Health LLC, Houston, TX

Adrian Ecker
Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Daniel Aeschbach
Division of Sleep and Human Factors Research, Institute of Aerospace Medicine, German Aerospace Center, Cologne, Germany

Alexander C Stahn
Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Christopher Jones
Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Kia Howard
Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Marc Kaizi-Lutu
Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

David F. Dinges
Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Haochung Shou
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

November 2021

NASA STI Program Report Series

The NASA STI Program collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- Help desk contact information:

https://www.sti.nasa.gov/sti-contact-form/ and select the "General" help request type.

Dynamic Ensemble Prediction of Cognitive Performance in Space

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

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

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

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

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

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

[*] Equal contribution

Abstract:

Astronauts are exposed to a unique set of stressors in spaceflight. Microgravity, isolation, confinement, and environmental and operational hazards: all of these can impact sleep, vigilant attention, and alertness, which are critical to mission success. In this paper, we seek to understand the most important predictors of alertness over the course of a space mission, using self-reported, cognitive, and environmental data collected from 24 astronauts on 6-month missions to the International Space Station (ISS). Alertness was repeatedly and objectively assessed on the ISS with a brief 3-minute Psychomotor Vigilance Test (PVT) that is highly sensitive to sleep deprivation. To relate PVT performance to time-varying and sparsely-measured environmental, operational, and psychological covariates, we propose an ensemble prediction model comprising of linear mixed effects regression, random forest, and functional concurrent regression models. An extensive cross-validation procedure reveals that this ensemble outperforms any one of its components alone. We also discover that a participant's past performance, reported fatigue and stress, and temperature and radiation exposure were among the most important variables associated with alertness. This method is broadly applicable to environmental studies where the main goal is accurate, individualized prediction involving a mixture of person-level traits and irregularly measured time series.

Keywords: cognitive performance, spaceflight, PVT, functional data analysis, ensemble

Introduction

During spaceflight, astronauts must perform cognitively intensive tasks that require sustained attention, often with disruptions to sleep and the circadian rhythm (Barger et al., 2014). Human performance deteriorates without proper rest, manifesting in longer reaction times and increased errors (Dinges, 1995), heightening the risk of accidents. As space travel is a costly and hazardous endeavor, it is critical to anticipate changes in alertness on a dynamic and individualized basis. While environmental and psychological correlates of alertness have been studied in the general population, highly trained and carefully selected astronauts are not necessarily represented in this population. In space, they are exposed to a unique set of conditions (Scully et al., 2019; Strangman et al., 2014; Thirsk et al., 2009), including microgravity, extended confinement and isolation, radiation exposure, and other environmental and operational extremes. The collective impact of these challenges on psychological health and cognitive performance are not yet fully understood (Roy-O'Reilly et al., 2021).

The goal of this study is to dynamically predict vigilant attention as a function of astronauts' past performance, self-reported stress and fatigue, demographic information, and variations in environmental stressors. Vigilant attention is assessed using the LRM-50 score (Basner et al., 2015), which is derived from a shortened version of the Psychomotor Vigilance Task (PVT) developed for spaceflight (Basner et al., 2011). The main challenge of predicting PVT performance is unraveling variation associated with the circadian rhythm, individual traits, psychological state, and the external environment (Olofsen et al., 2010; Rupp et al., 2012). Previously, PVT performance was successfully predicted via a two-process model (Borbély, 2008), which incorporates a system of differential equations to describe homeostatic and circadian pressures governing sleep. While such models have expanded our understanding of the biological mechanisms of sleep and have been greatly adapted (McCauley et al., 2013; Postnova et al., 2018; St. Hilaire et al., 2007), they are often deterministic and so preclude statistical comparisons; even models with person-level random effects (Van Dongen et al., 2007) 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 (Bhat et al., 2018; Graw et al., 2004), 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 (Jewett et al., 1999). Methods which have addressed these have mainly considered mixed-effect models (Bermudez et al., 2016) or an ensemble of mixed-effects and random forest models (Cochrane et al., 2021). However, neither of these methods can explicitly model time-varying predictors whose effects themselves are time-varying, as in the case of circadian effects (Blatter & Cajochen, 2007) or acclimation (Liu et al., 2016; Williams et al., 2009).

In this paper, we propose a 3-model ensemble prediction scheme consisting of a linear mixed effects model (Rencher & Schaalje, 2007), a random forest model (Breiman, 2001), and a functional concurrent model (Leroux et al., 2018), 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 (McCauley et al., 2013) as a covariate, with the aim to connect the biomathematical and statistical models commonly used to predict performance. Similar to previous methods (Cochrane et al., 2021), we employ forward-chaining cross-validation (Bergmeir & Benítez, 2012) to demonstrate that the ensemble predicts best over the entire mission compared to any single component alone. The remainder of this paper is organized as follows: in section 2, we introduce the Reaction Self-Test (RST) and environmental data collected from astronauts and instruments aboard the International Space Station (ISS). In section 3, we describe a variable selection and model validation procedure that emphasizes prediction accuracy. In section 4, we consider the best-performing ensemble model and identify key predictors of alertness. In section 5, we conclude with a discussion of our findings and future work.

Data

Participants and Protocol

Reaction Self-Test (i.e., self-report and PVT-B data, see below) data was collected from n = 24 astronauts over 19 mission increments between 2009 and 2013 (Table 1). Astronauts spent an average of 160 (SD = 19) days, with a range of 123 to 192 days on the ISS. Ahead of the mission, astronauts were scheduled to complete the RST once 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. Whilst in-flight during the mission, astronauts were instructed to complete the RST twice a day (upon waking and in the 2 hours before sleep) every four 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 n = 2,968 RST observations. The original study as well as this re-analysis were approved by the IRB of Johnson Space Center and the University of Pennsylvania (for data analysis). Participants provided written informed consent prior to study participation and re-consented for the re-analysis of the data.

Reaction Self-Test (RST)

The RST consists of a brief survey followed by a validated, brief (3-minute) version of PVT (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 (Basner & Dinges, 2011). Astronauts were instructed to monitor a box on the laptop screen and press the space bar once a millisecond counter appeared in the box and started incrementing. The reaction time was then displayed

for one second. Participants were instructed to be as fast as possible without hitting the spacebar in the absence of a stimulus. The PVT is a sensitive measure of vigilant attention, and has been well-established as a tool to detect acute and chronic sleep deprivation and circadian misalignment, conditions highly prevalent in spaceflight (Barger et al., 2014). The PVT has negligible aptitude and learning effects (Basner et al., 2018), and is ecologically relevant as sustained attention deficits and slow reactions affect many real-world tasks (e.g., operating a moving vehicle) (Dinges, 1995).

The survey includes a sleep diary, 11-point rating scales on tiredness, mental fatigue, physical exhaustion, stress, sleepiness, with the final rating depending on the time of day: workload (evening administration only) or sleep quality (morning administration only). During both the morning and evening RSTs astronauts were asked to list the name, dose unit, and doses taken of all medications they ingested before going to bed last night (morning RST) and since awakening in the morning (evening RST). Additionally, in the evening RST they were asked to list caffeinated food or beverages they had taken 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. 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 performance metrics, we derive 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 (Basner et al., 2015). LRM-50 is a likelihood ratio-based metric that assigns a likelihood of being sleep-deprived depending on the response time to each PVT stimulus; high values of the LRM-50 translate to low levels of alertness. Although the number of lapses in attention is a more commonly used PVT metric, it is less able to differentiate high performers, such as astronauts, compared to LRM-50. The standardized LRM-50 outcome, which is formed by scaling each individual's LRM-50 by their person-level mean and standard deviation, was considered in our analyses but not shown to drastically improve prediction accuracy.

Environmental Data

During the study period, five domains of environmental measures were also recorded from the ISS. Radiation exposure levels were obtained from the Space Radiation Analysis Group at Johnson Space Center, and were summarized in daily dosage units (mGy) based on active dosimeters located on the interior and exterior of the space station. The radiation dose was defined as the sum of radiation due to Galactic Cosmic Rays (GCR) and the South Atlantic Anomaly (SAA). Next, oxygen (O_2) and carbon dioxide (CO_2) levels in units of mmHg were collected from Major Constituent Analyzer (MCA) sensors located in the space station's air circulation system. Finally, temperature in Celsius and noise level in A-weighted decibels (dBA) were recorded by sensors distributed across the ISS. Environmental variables were only collected during the in-flight period but not during pre- or post-flight periods.

Demographic and Operational Data

Demographic information included sex, age at ISS docking, nationality, space agency, military background, educational attainment, marital status, number of children, 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; proximity of test to dock maneuvers, undock maneuvers, or EVAs; and the scheduled sleep period. The latter was used to identify sleep shifts.

Methods

Derived Predictors

We derived several predictors based on the RST data: a fatigue composite score, four medication flags, and predicted PVT lapses. The fatigue composite score was created based on principal components analysis (PCA) of the 11-point scales on which the crew rated several behavioral states (see above) before taking the PVT. The score was calculated as the weighted average of the VAS questions, with weights determined by the loadings of the first PCA component. Table S1 shows the loadings of the first PC, which accounted for 48.3% of the variance. Higher values of the fatigue composite variable correspond to higher workload, increased tiredness, more stress, and worse sleep quality. Next, medication use was coded as a binary variable for four broad categories: pain medications, sleep aids, decongestants, and antihistamines. These categories were chosen as their use may affect sleep or be correlated with conditions affecting sleep (Marin et al., 2006; Meltzer, 1990; Tannenbaum et al., 2012). The final derived covariate was the number of predicted PVT lapses under a two-factor model (McCauley et al., 2013) given the sleep schedule collected in the RST.

Data Integration and Interpolation

To integrate the environmental data with the RST data, several strategies were required. This is because different variables were recorded at different time intervals: RST was collected twice a day every few days; radiation exposure and other operational variables were measured daily; temperature, noise level, CO_2, and O_2 were measured multiple times per day or minute. For each RST observation, the value of radiation (and other day-level variables) from that day was used. For more finely-measured variables, we used the hourly average from the hour that the RST was taken, if available.

When the daily or hourly value of an environmental variable was unavailable, we used two interpolation strategies. Temperature, O_2, and CO_2 data had a relatively low rate of missingness, so the locally estimated scatterplot smoothing (LOESS, neighborhood parameter α = 1) value was used if the hourly average was unavailable. The expected PVT lapses could not be predicted for periods where no sleep diary information was provided, and in these periods the number of lapses was also interpolated using LOESS. Noise levels were only recorded on a handful of days: for this variable, the average noise during daytime (7:00 AM to 10:59 PM UTC)

and nighttime (11:00 PM to 6:59 AM UTC) hours was interpolated separately using linear interpolation. The distribution of smoothed and unsmoothed environmental data is shown in Supporting Table S2.

Finally, temperature was measured using multiple sensors across the ISS, including the US Lab and the Node 2 modules. Therefore, we also performed spatial matching and aggregation to achieve the best estimate for each RST. When the RST was taken in Node 2 or the US Lab, only the temperature 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 empirical frequency of RSTs taken in these modules.

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. 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 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. 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 $X_{it} = (X_{it}^{(1)}, \dots, X_{it}^{(p)})$, person-specific random intercept b_i, and noise 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}.

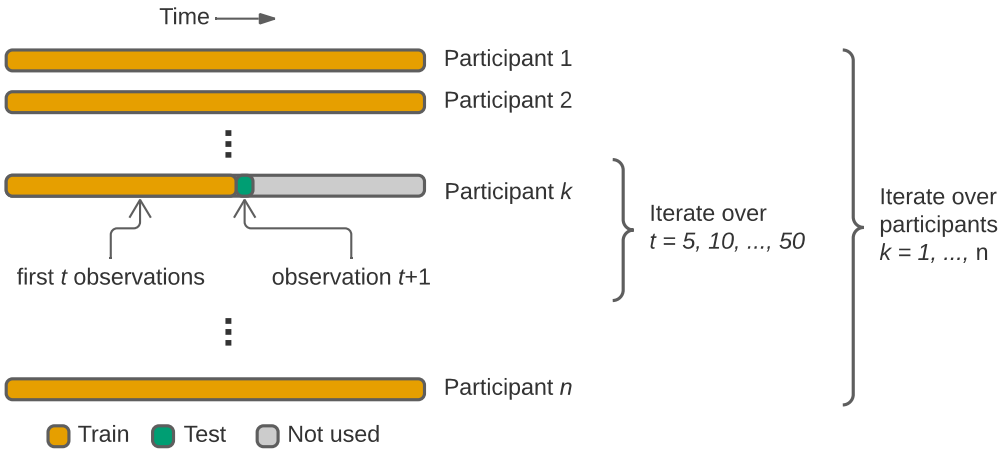
By contrast, the random forest model (Breiman, 2001) specifies no closed form for the relationship between y_{it} and X_{it}; rather, it uses an aggregate of decision trees to identify splitting points for continuous variables that optimally predict the outcome. While prone to overfitting, random forests 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 (Leroux et al., 2018): for participant i and time t_{ij}, the time-varying outcomes y_{ij} are related to p covariates $X_{ij}^{(1)}, \dots, X_{ij}^{(p)}$ through the following:

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

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. Finally, the ensemble prediction was constructed as the unweighted average of predictions from the LME, random forest, and functional concurrent model. All data analysis was performed using R version 3.6.1 (R Core Team, 2019), 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 (Figure 1). For participant k and training length $t \in \{5, 10, \ldots, 45, 50\}$, a given model was fit on the first t RST observations from participant k and all data from all other participants. The squared prediction error at day $t + 1$ was then averaged over all training lengths t for participant k to obtain a person-level mean squared error (MSE). These were again averaged over all participants to obtain an overall MSE for each model.

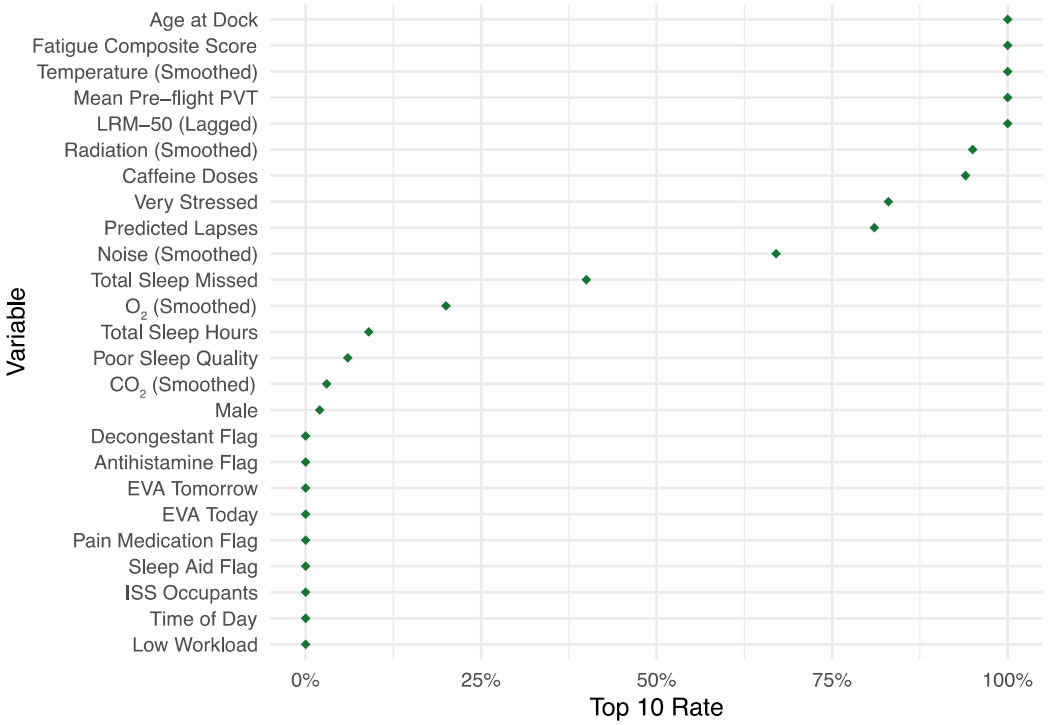
Other models, such as multivariate linear regression, time series regression (using the *dyn* R package), and generalized additive models (using the *gamlss* R package) were considered at this stage, but did not perform well to justify inclusion in the final ensemble.

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 mean square error (MSE) (%IncMSE) when permuting that variable within a random forest model. By repeating this process 100 times and sampling 50% of the data each time, 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. In statistical analyses, we considered both models fit on the full set of predictors, as well as using the subset of the most important variables.

Results

Based on the random forest importance ranking, the most important predictors of LRM-50 included individual characteristics such as age and average pre-flight PVT score; the most recent LRM-50 score; psychological factors such as the fatigue score and feelings of stress; sleep-related factors including caffeine intake, predicted lapses under a two-factor model, and the amount of sleep missed (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 getting up); and smoothed environmental measurements including temperature and radiation exposure (Figure 2). Comparatively, we found that sex, medication use, scheduled EVAs, and workload were less important to predicting LRM-50.



Our experiments indicate that the ensemble model performed better than any single model alone over various training lengths t after forward-chaining cross-validation (Figure 3).

Interestingly, in the testing data, the MSE of all models decreased with longer training period t when t was small, but then increased with t for larger t. One potential explanation is the lack of testing data for higher values of t. We also found that the linear mixed effects model performed better than functional concurrent and random forest with smaller t, while the other two models outperformed LME for higher t. Together, these suggest that the random forest and functional models are more suitable for longer time series, while the LME performs sufficiently for shorter time series when fewer dynamic relationships are observed between variables.

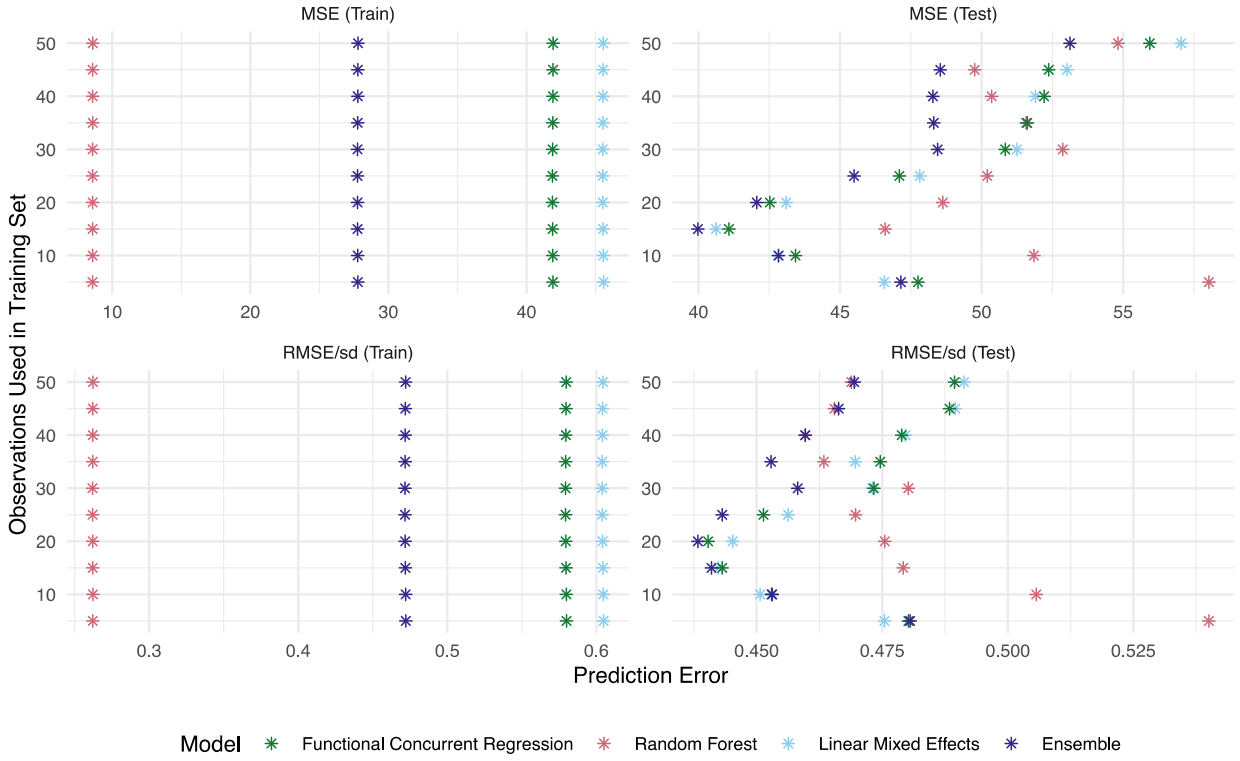


Figure 3. Prediction accuracy among each of the component models and the ensemble. Model performance was measured using the mean squared error (MSE) in predicting LRM-50. A standardized measure of prediction error, which can be used to compare across different outcomes, was obtained by dividing the root MSE (RMSE) by the standard deviation of the outcome in the training set. In other words, the RMSE/sd represents the variation in the observed LRM-50 data that is not explained by the model.

We next considered the effects of environmental conditions on LRM-50 by examining the corresponding LME coefficients (Table S3) and functional concurrent regression heat maps (Figure 4). In both models, we found that better performance (i.e., lower LRM-50) was associated with lower radiation exposure and higher CO_2 levels. The functional model further revealed that the negative effects of radiation exposure and positive effects of CO_2 were attenuated over time, suggesting acclimation. Finally, the functional model identified non-monotonic relationships between performance and temperature and oxygen, suggesting the existence of an optimal range for these measures; this type of relationship could not be assessed by the mixed model.

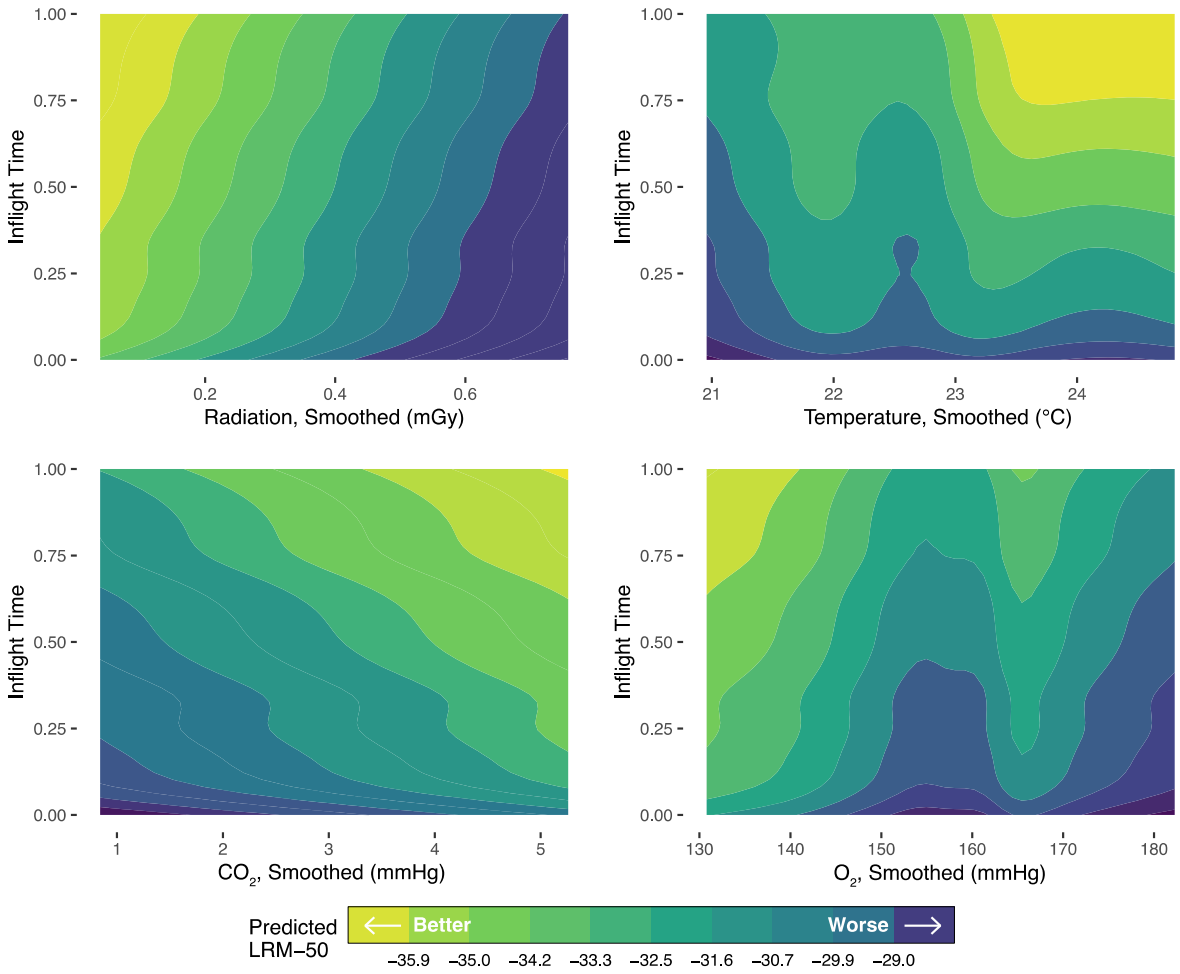


Figure 4. Using predictions from the functional concurrent model, these heat maps show how the non-linear effects of environmental variables on LRM-50 vary over time in mission (ranging from 0 to 1 representing the proportion of mission time elapsed). The same functional concurrent model, which was fit on the entire observed data, was used for each panel. For each environmental variable, predictions were made at a regular grid of timepoints 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. We find that better performance (i.e., lower LRM-50, indicated by lighter yellow regions) is generally associated with lower radiation, moderate to higher temperatures, higher CO_2, and moderate and lower O_2.

The ensemble model was implemented as a user-friendly and interactive R Shiny application (Figure 5). Given the data and the fitted ensemble model, the application displays individualized LRM-50 predictions and other model diagnostic information. To encourage hypothesis generation, the value of predictor variables can also be "toggled," allowing the user to view the predicted LRM-50 under hypothetical sets of conditions. Finally, the application includes each participant's entire trajectory of predicted and observed LRM-50 scores as a function of the training period.

An astronaut's LRM-50 score can be predicted at an arbitrary number of future timepoints, but this requires knowledge of environmental conditions and other covariates at those timepoints.

In practice, we may obtain the best prediction at a particular timepoint 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 timeline of observed and predicted performance for each participant (Figure 6). The root mean squared error (RMSE) of these "chained" predictions over time ranged from 4.06 to 11.98 among astronauts.

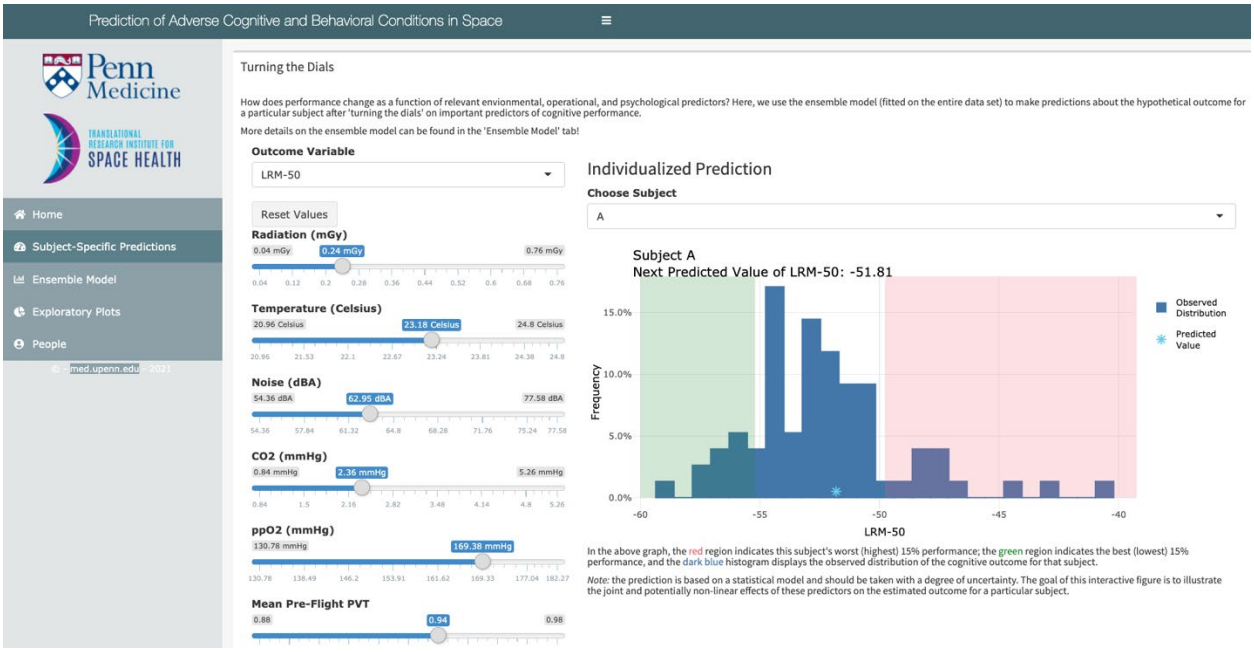


Figure 5. A screenshot of the R Shiny application implementing the ensemble prediction model. In the left panel, the user may "toggle" the value of each predictor (pre-set to averages observed for the individual astronaut). In the right panel, the individualized predicted LRM-50 score for the selected participant is displayed (blue star at bottom of graph), along with the distribution of that astronaut's observed scores; the red and green regions correspond to that astronaut's worst 10% and best 10% scores.

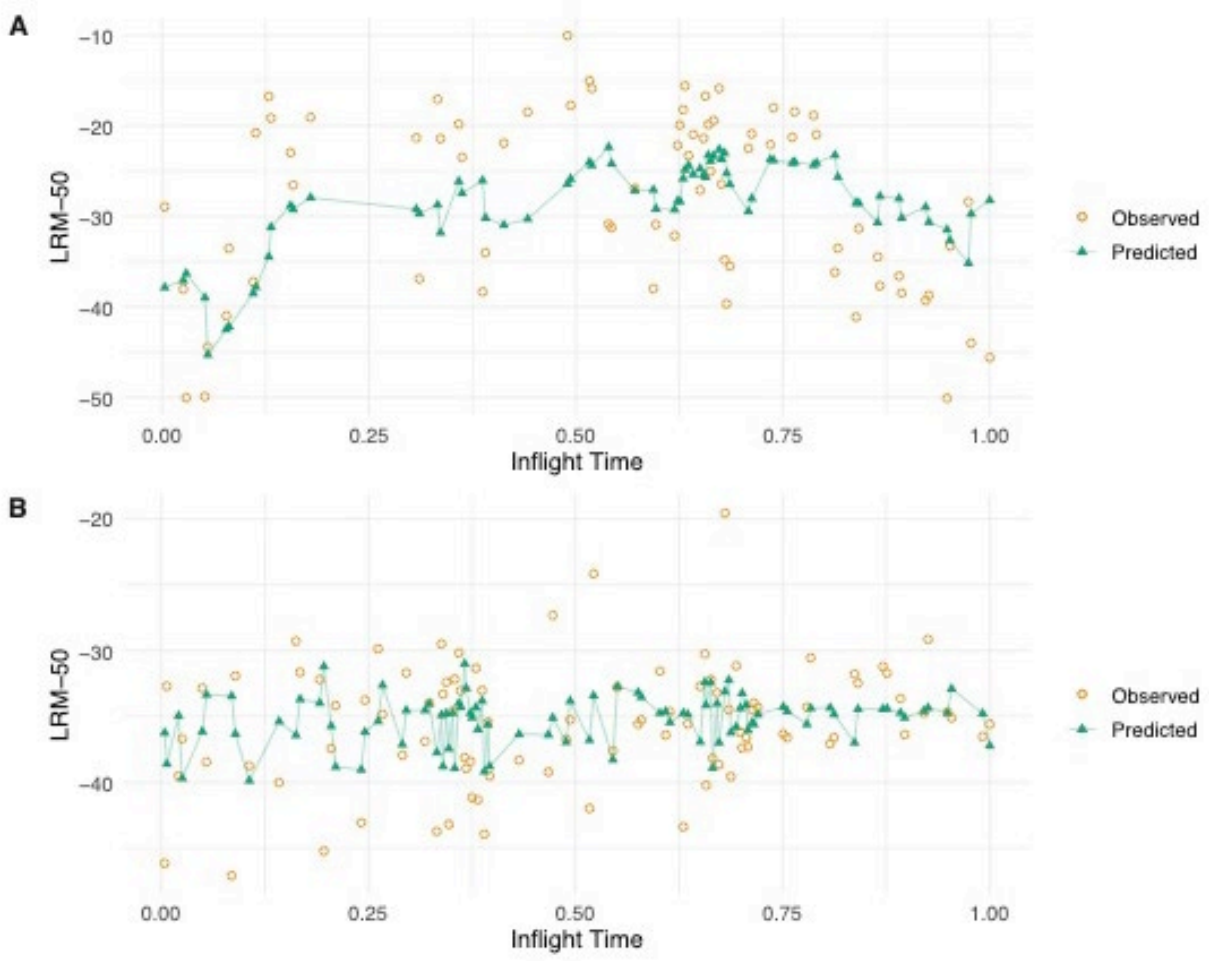


Figure 6. At each timepoint and for a given astronaut, LRM-50 can be predicted by fitting the model on all preceding data from that astronaut and the full data from other astronauts. The prediction (solid green triangles) is then made using covariate values from that timepoint. The actual values of LRM-50 are displayed as hollow yellow circles. (A) A participant with prediction MSE in the highest (worst) 25th percentile. (B) A participant with prediction MSE in the lowest (best) 25th percentile.

Discussion

In this paper, we proposed an ensemble model to predict cognitive performance in astronauts during spaceflight missions. In contrast to previous methods that employed a single prediction method (Bermudez et al., 2016) or ensemble (Cochrane et al., 2021), we added a dynamic component to model time-varying covariate effects using a functional concurrent model (Leroux et al., 2018). The resulting model can flexibly and accurately predict LRM-50 alertness score. We also identified the most important predictors of alertness as a combination of individual traits, dynamic psychological state, and environmental conditions.

Spaceflight is a complex environment with myriad psychological, operational and environmental stressors that change throughout the course of a mission. Behavioral health studies in space or ground-based space analog environments have traditionally focused on a small number of stressors (Basner et al., 2021; Connaboy et al., 2020; Strangman et al., 2014).

While these studies provide important mechanistic information, they fail to address the intertwined and time-varying effects of several concurrent stressors. 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 on vigilant attention, that also includes individualized predictions of future performance.

Ensemble models of machine learning models are increasingly popular in human health studies due to their flexibility and ability to accommodate non-standard data types (Rose, 2018). 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 (Leroux et al., 2018) are able to powerfully capture dynamic effects. Furthermore, the incorporation of models with both scalar and functional random effects are useful to making individualized predictions (Figure 6).

Our findings have three main applications:

1) The models were used to identify the most relevant predictors of psychomotor vigilance in spaceflight. Self-assessments of fatigue and stress, temperature and radiation exposure, caffeine consumption, and past performance were identified as some of the most important correlates of performance. This variable selection helps space agencies like NASA to concentrate research and mitigation measures on those variables. Once new data is available from future studies, the ensemble model could easily expand to include other predictors of interest.
2) Relationships between two predictor variables can be visualized in the R shiny application (Figure 5) and in heatmaps (Figure 4). This tool therefore facilitates the generation of hypotheses that can later be empirically verified.
3) Exploration-class space missions will involve communication delays and increase crew autonomy. Self-administered tests that assess readiness-to-perform can therefore be a helpful tool in guiding astronaut operational decisions. One application of the ensemble model is real-time decision-making aboard the ISS. The predicted LRM-50, given the most current environmental and RST data, could be used as a score of astronaut readiness ahead of mission critical tasks (e.g., 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 cognitive 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 has several limitations. Because the main goal of ensemble prediction is accuracy, it is not generally possible to make inferences about 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. Secondly, our ensemble

prediction weighted each model equally, as no model consistently over- or under-performed. Future work could use cross-validation to determine the weights empirically. Finally, 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 (Smith et al., 2009).

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 cognitive performance and sleep in space, has identified promising avenues in modelling dynamic and personalized profiles of alertness. Such tools could have important implications for safety and decision-making in one of the world's most high-profile and dangerous occupations.

Acknowledgements

This work is supported by the Translational Research Institute for Space Health through NASA Cooperative Agreement NNX16AO69A. 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].

Tables

Table 1. Summary characteristics of the astronauts with RST data. Table values are mean (standard deviation) and count (percent) for continuous and categorical variables, respectively.

	Overall
	(N = 24)
Sex, n (%)	
Male	19 (79.2)
Female	5 (20.8)
Age at Dock	48.2 (4.78)
Agency, n (%)	
NASA	16 (66.7)
Non-NASA	8 (33.3)
Average Pre-Flight PVT Score	0.95 (0.02)
Inflight RST Observations	87.2 (18.8)

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

		Overall
		(N = 2968)
	Period, n (%)	
	Pre-Flight	506 (17.0)
	Inflight	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 VAS	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)
EVA	EVA Today Flag, n (%)	8 (0.27)
	EVA Tomorrow Flag, n (%)	23 (0.77)

Supporting Material

Table S1. Principal components analysis (PCA) of the Self-Report VAS variables and the loadings on the first principal component are shown below. These were then used as weights to calculate a composite variable called the fatigue index. Based on these weights, higher values of the fatigue index correspond to greater workloads, more stress, and more physical and mental tiredness.

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 S2. All predictors used in the prediction models.

Source	*Category*	*Variables*	*Data Type*
Reaction Self-Test (RST)	PVT Performance	LRM-50	Time-Varying
		Standardized LRM-50	Time-Varying
	RST Type	Time of Day (Morning/Evening)	Time-Varying
	Sleep	Time in Bed Sleeping	Time-Varying
		Time in Bed Not Sleeping	Time-Varying
		Predicted Lapses	Time-Varying
	Self-Report	Very Stressed	Time-Varying
		Low Workload	Time-Varying
		Poor Sleep Quality	Time-Varying
		Fatigue Composite Score	Time-Varying
	Medications	Caffeine Doses	Time-Varying
		Sleep Aid Flag	Time-Varying
		Decongestant Flag	Time-Varying
		Antihistamine Flag	Time-Varying
		Pan Medication Flag	Time-Varying
Demographics		Age at Dock	Scalar
		Sex	Scalar
		Average Pre-Flight PVT Score	Scalar
Environmental		Radiation	Time-Varying
		Temperature	Time-Varying
		Noise	Time-Varying
		CO_2	Time-Varying
		O_2	Time-Varying
		ISS Occupancy	Time-Varying

Table S3. 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).

Variable	Coefficient
(Intercept)	-0.314
Radiation, Smoothed (mGy)	0.010
Noise, Smoothed (dBA)	-0.059
CO_2, Smoothed (mmHg)	-0.012
O_2, Smoothed (mmHg)	-0.006
ISS Occupants	0.018
Temperature, Smoothed (°C)	-0.060
Sex = Male	0.302
Age at Dock	-0.071
Average Pre-flight PVT Score	-0.201
Sleep Aid Flag	0.056
Antihistamine Flag	0.078
Time of Day = Morning Test	0.123
Fatigue Composite Score	0.181
Low Workload	-0.016
Poor Sleep Quality	-0.005
Very Stressed	-0.013
Total Sleep Hours	-0.061
Total Sleep Missed	-0.005
Predicted Lapses	0.006
Caffeine Doses	0.022
LRM-50 (Lagged)	0.122

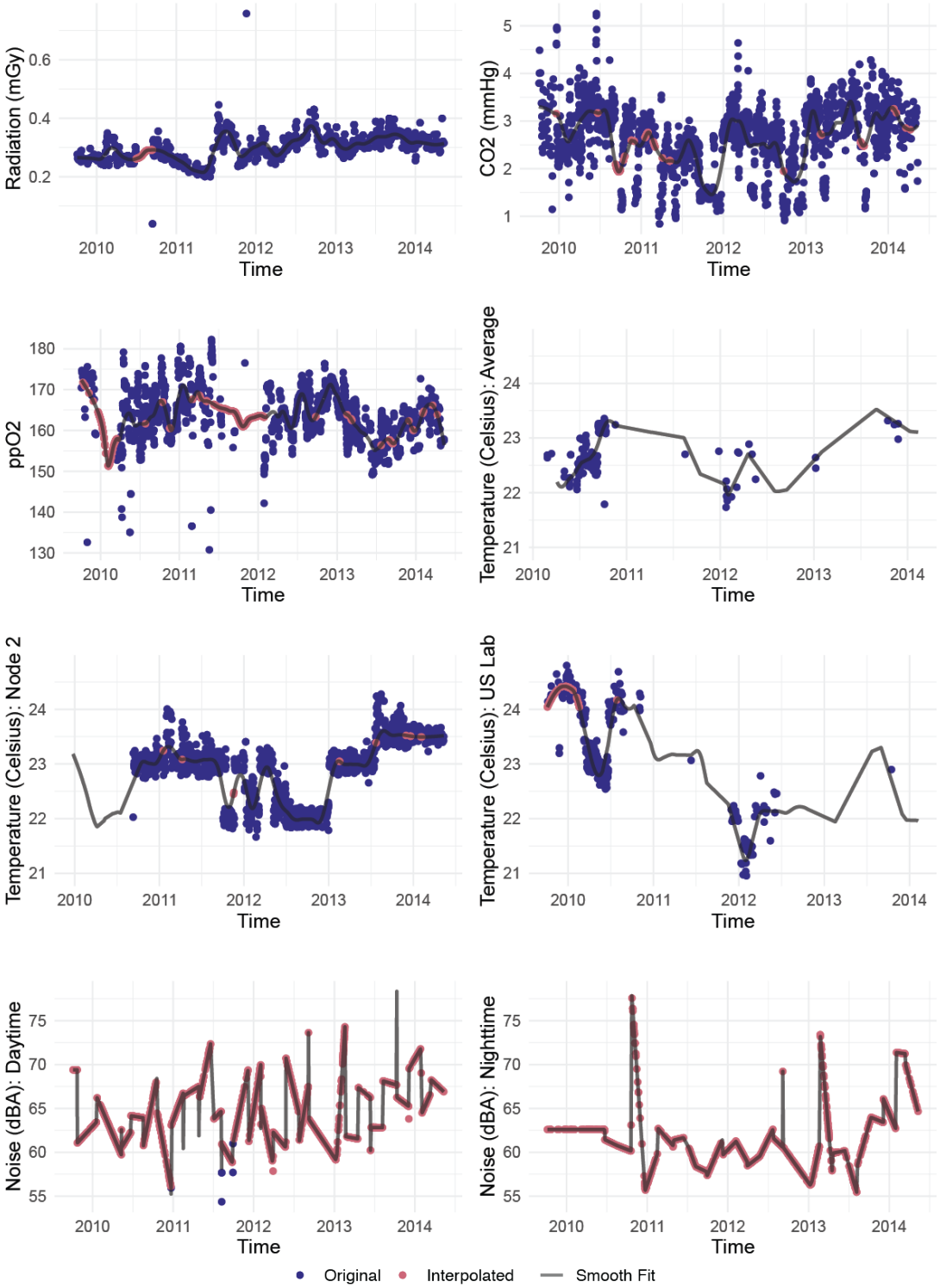


Figure S1. For each RST observation, the corresponding value of the environmental variable was found by using the observed value (if available) or the interpolated value formed by neighboring observations. These plots illustrate the LOESS curves (black line) fit to the entire environmental data for radiation, temperature (separately for each location), CO_2, and O_2. A linear interpolation was used for noise (separately for daytime and nighttime). Each blue dot corresponds to the observed hourly average (CO_2, O_2, temperature, noise) or daily average (radiation) that was used for an RST observation; the red dot indicates that the interpolated value was used.

References

Barger, L., Flynn-Evans, E., A, W. K., Ronda, J., Wang, W., Wright, K., & Czeisler, C. (2014). Prevalence of Sleep Deficiency and Hypnotic Use Among Astronauts Before, During and After Spaceflight: An Observational Study. *Aviation, Space, and Environmental Medicine*, *13*(9), 904–912. https://doi.org/10.1016/S1474-4422(14)70122-X.Prevalence

Basner, M., & Dinges, D. F. (2011). Maximizing Sensitivity of the Psychomotor Vigilance Test (PVT) to Sleep Loss. *Sleep*, *34*(5), 581–591. https://doi.org/10.1093/sleep/34.5.581

Basner, M., Hermosillo, E. B. A., Nasrini, J. B. S., McGuire, S., Saxena, S., Moore, T. M., Gur, R. C., & Dinges, D. F. (2018). Repeated administration effects on psychomotor vigilance test performance. *Sleep*, *41*(1). https://doi.org/10.1093/sleep/zsx187

Basner, M., Mcguire, S., Goel, N., Rao, H., & Dinges, D. F. (2015). A new likelihood ratio metric for the psychomotor vigilance test and its sensitivity to sleep loss. *Journal of Sleep Research*, *24*(6), 702–713. https://doi.org/10.1111/jsr.12322

Basner, M., Mollicone, D., & Dinges, D. F. (2011). Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta Astronautica*, *69*(11–12), 949–959. https://doi.org/10.1016/j.actaastro.2011.07.015

Basner, M., Stahn, A. C., Nasrini, J., Dinges, D. F., Moore, T. M., Gur, R. C., Mühl, C., Macias, B. R., & Laurie, S. S. (2021). Effects of head-down tilt bed rest plus elevated CO 2 on cognitive performance. *Journal of Applied Physiology*, *130*(4), 1235–1246. https://doi.org/10.1152/japplphysiol.00865.2020

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Bermudez, E. B., Klerman, E. B., Czeisler, C. A., Cohen, D. A., Wyatt, J. K., & Phillips, A. J. K. (2016). Prediction of vigilant attention and cognitive performance using self-reported alertness, circadian phase, hours since awakening, and accumulated sleep loss. *PLoS ONE*, *11*(3), 1–18. https://doi.org/10.1371/journal.pone.0151770

Bhat, S., Gupta, D., Akel, O., Polos, P. G., DeBari, V. A., Akhtar, S., McIntyre, A., Ming, S. X., Upadhyay, H., & Chokroverty, S. (2018). 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 Medicine*, *49*, 81–89. https://doi.org/10.1016/j.sleep.2018.06.012

Blatter, K., & Cajochen, C. (2007). Circadian rhythms in cognitive performance: Methodological constraints, protocols, theoretical underpinnings. *Physiology and Behavior*, *90*(2–3), 196–208. https://doi.org/10.1016/j.physbeh.2006.09.009

Borbély, A. A. (2008). Two-Process Model of Sleep Regulation. *Encyclopedia of Neuroscience*, 4146–4146. https://doi.org/10.1007/978-3-540-29678-2_6166

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cochrane, C., Ba, D., Klerman, E. B., & St. Hilaire, M. A. (2021). An ensemble mixed effects model of sleep loss and performance. *Journal of Theoretical Biology*, *509*, 110497. https://doi.org/10.1016/j.jtbi.2020.110497

Connaboy, C., Sinnott, A. M., LaGoy, A. D., Krajewski, K. T., Johnson, C. D., Pepping, G.-J., Simpson, R. J., Bower, J. L., & Alfano, C. A. (2020). Cognitive performance during prolonged

periods in isolated, confined, and extreme environments. *Acta Astronautica*, *177*, 545–551. https://doi.org/10.1016/j.actaastro.2020.08.018

Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, *4*, 4–14. https://doi.org/10.1111/j.1365-2869.1995.tb00220.x

Graw, P., Kräuchi, K., Knoblauch, V., Wirz-Justice, A., & Cajochen, C. (2004). Circadian and wake-dependent modulation of fastest and slowest reaction times during the psychomotor vigilance task. *Physiology and Behavior*, *80*(5), 695–701. https://doi.org/10.1016/j.physbeh.2003.12.004

Jewett, M. E., Dijk, D. J., Kronauer, R. E., & Dinges, D. F. (1999). Dose-response relationship between sleep duration and human psychomotor vigilance and subjective alertness. *Sleep*, *22*(2), 171–179. https://doi.org/10.1093/sleep/22.2.171

Leroux, A., Xiao, L., Crainiceanu, C., & Checkley, W. (2018). Dynamic prediction in functional concurrent regression with an application to child growth. *Statistics in Medicine*, *37*(8), 1376–1388. https://doi.org/10.1002/sim.7582

Liu, Q., Zhou, R. L., Zhao, X., Chen, X. P., & Chen, S. G. (2016). Acclimation during space flight: Effects on human emotion. *Military Medical Research*, *3*(1), 3–7. https://doi.org/10.1186/S40779-016-0084-3

Marin, R., Cyhan, T., & Miklos, W. (2006). Sleep Disturbance in Patients With Chronic Low Back Pain. *American Journal of Physical Medicine & Rehabilitation*, *85*(5), 430–435. https://doi.org/10.1097/01.phm.0000214259.06380.79

McCauley, P., Kalachev, L. V., Mollicone, D. J., Banks, S., Dinges, D. F., & Van Dongen, H. P. A. (2013). Dynamic circadian modulation in a biomathematical model for the effects of sleep and sleep loss on waking neurobehavioral performance. *Sleep*, *36*(12), 1987–1997. https://doi.org/10.5665/sleep.3246

Meltzer, E. O. (1990). Antihistamine- and Decongestant-Induced Performance Decrements. *Journal of Occupational and Environmental Medicine*, *32*(4), 327–334. https://doi.org/10.1097/00043764-199004000-00013

Olofsen, E., Dongen, H. P. a Van, Mott, C. G., Balkin, T. J., Terman, D., Reed, W., & Spring, S. (2010). Current Approaches and Challenges to Development of an Individualized Sleep and Performance Prediction Model. *Sleep (Rochester)*, 24–43.

Postnova, S., Lockley, S. W., & Robinson, P. A. (2018). Prediction of Cognitive Performance and Subjective Sleepiness Using a Model of Arousal Dynamics. *Journal of Biological Rhythms*, *33*(2), 203–218. https://doi.org/10.1177/0748730418758454

Rencher, A. C., & Schaalje, G. B. (2007). Linear Models in Statistics. In *Linear Models in Statistics*. https://doi.org/10.1002/9780470192610

Rose, S. (2018). Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open*, *1*(4), e181404. https://doi.org/10.1001/jamanetworkopen.2018.1404

Roy-O'Reilly, M., Mulavara, A., & Williams, T. (2021). A review of alterations to the brain during spaceflight and the potential relevance to crew in long-duration space exploration. *Npj Microgravity*, *7*(1), 1–9. https://doi.org/10.1038/s41526-021-00133-z

Rupp, T. L., Wesensten, N. J., & Balkin, T. J. (2012). Trait-Like Vulnerability to Total and Partial Sleep Loss. *Sleep*, *35*(8), 1163–1172. https://doi.org/10.5665/sleep.2010

Scully, R. R., Basner, M., Nasrini, J., Lam, C. wing, Hermosillo, E., Gur, R. C., Moore, T., Alexander, D. J., Satish, U., & Ryder, V. E. (2019). Effects of acute exposures to carbon

dioxide on decision making and cognition in astronaut-like subjects. *Npj Microgravity*, *5*(1). https://doi.org/10.1038/s41526-019-0071-6

Smith, A. D., Genz, A., Freiberger, D. M., Belenky, G., & Van Dongen, H. P. A. (2009). *Chapter 8 Efficient Computation of Confidence Intervals for Bayesian Model Predictions Based on Multidimensional Parameter Space* (pp. 213–231). https://doi.org/10.1016/S0076-6879(08)03808-1

St. Hilaire, M. A., Klerman, E. B., Khalsa, S. B. S., Wright, K. P., Czeisler, C. A., & Kronauer, R. E. (2007). Addition of a non-photic component to a light-based mathematical model of the human circadian pacemaker. *Journal of Theoretical Biology*, *247*(4), 583–599. https://doi.org/10.1016/j.jtbi.2007.04.001

Strangman, G. E., Sipes, W., & Beven, G. (2014). Human cognitive performance in spaceflight and analogue environments. *Aviation Space and Environmental Medicine*, *85*(10), 1033–1048. https://doi.org/10.3357/ASEM.3961.2014

Tannenbaum, C., Paquette, A., Hilmer, S., Holroyd-Leduc, J., & Carnahan, R. (2012). A Systematic Review of Amnestic and Non-Amnestic Mild Cognitive Impairment Induced by Anticholinergic, Antihistamine, GABAergic and Opioid Drugs. *Drugs & Aging*, *29*(8), 639–658. https://doi.org/10.1007/BF03262280

Thirsk, R., Kuipers, A., Mukai, C., & Williams, D. (2009). The space-flight environment: the International Space Station and beyond. *Canadian Medical Association Journal*, *180*(12), 1216–1220. https://doi.org/10.1503/cmaj.081125

Van Dongen, H. P. A., Mott, C. G., Huang, J. K., Mollicone, D. J., McKenzie, F. D., & Dinges, D. F. (2007). 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*(9), 1129–1143. https://doi.org/10.1093/sleep/30.9.1129

Williams, D., Kuipers, A., Mukai, C., & Thirsk, R. (2009). Acclimation during space flight: Effects on human physiology. *Cmaj*, *180*(13), 1317–1323. https://doi.org/10.1503/cmaj.090628