Machine Learning-Ready Data Sets for the Analysis and Nowcasting of Atmospheric Radiation at Aviation Altitudes

V. M. Sadykov[1], Z. M. Watkins[1], D. Kempton[1], W. Jones[1], S. K C[1], G. T. Goodwin[1], X. He[1], W. K. Tobiska[2], I. Kitiashvili[3], C. Mertens[4], S. Ranjan[3], D. G. Deardorff[5], R. Spaulding[5], S. L. Pokrandt[5]

[1]Georgia State University
[2]Space Environment Technologies
[3]NASA Ames Research Center
[4]NASA Langley Research Center
[5]InuTeq, LLC

Key Points:

- We present an open-access machine learning-ready dataset for nowcasting the atmospheric radiation environment based on a diverse set of Geospace properties.
- Flight data points are split into three partitions with non-overlapping flights that equally sample the parameter space.
- The nowcasting test case demonstrates results comparable to predictions of physics-based radiation environment models.

Corresponding author: Viacheslav M Sadykov, vsadykov@gsu.edu

Corresponding author: Zachary Mark Watkins, zwatkins1@student.gsu.edu

Abstract

Nowcasting and forecasting of the radiation environment in the Earth's lower atmosphere are critical for the safety of aircraft and spacecraft crews and passengers. Currently, this problem is addressed by employing physics-based particle transport and precipitation models. However, given the increased number of radiation measurements available to the community, it is possible to start developing data-driven approaches. We prepared Machine Learning-ready (ML-ready) datasets to nowcast the effective dose rates at aviation altitudes. The presented datasets contain 92,476 individual measurements from 589 flights obtained by the Automated Radiation Measurements for Aerospace Safety (ARMAS) experiment from 2013 to 2023. The ARMAS measurements are augmented with the properties of the Geospace environment, such as solar soft X-ray and proton fluxes, solar wind properties, secondary cosmic ray neutrons, space weather indexes, and global solar activity indicators (such as daily sunspot number). ARMAS data are separated into three partitions, ensuring that (1) the data points from a single flight remain within the same partition, and (2) each partition samples the flight locations and Geospace environment conditions equally. Several versions of the datasets allow predictions based on point-in-time measurements and use up to 24 hours of Geospace parameter history. The test of the use case demonstrates a possibility of nowcasting ARMAS measurements with accuracies slightly better than the considered physics-based models. The publicly available ML-ready datasets could serve as the first step in data preparation for ML-driven nowcasting and forecasting of the radiation environment.

Plain Language Summary

Understanding the radiation levels in the Earth's atmosphere and predicting them at the locations along flight routes is important for the aviation industry. Machine Learning (ML) techniques are often used nowadays for prediction tasks. However, an extensive data preparation phase is required before applying an ML algorithm for predictions. Given the advantage of increasing volumes of the radiation measurement data, we present ML-ready datasets that allow users to bypass the data preparation stage and go directly to the phase of testing ML models for aviation radiation prediction. The ML-ready dataset is publicly available via the Radiation Data Portal[1].

1 Introduction

Understanding the radiation environment at the aviation altitudes (approximately $8 - 17$ km) remains an important and not fully explored topic for the aviation community. Integrating the effective dose rate along the flight trajectory constitutes the total effective dose absorbed by the aircraft crew and passengers. Given that there are recommended limits for the annual effective dose (20 mSv/yr and 1 mSv/yr for 5 years for the radiation occupation workers and for the general public, correspondingly; Cho et al., 2017), accurate monitoring and prediction of the radiation environment becomes essential.

Galactic cosmic rays cascading in Earth's atmosphere are the primary contributors to effective dose rates. In addition, the atmospheric radiation dose rates can increase during strong solar energetic particle (SEP) events (Reames, 2021). For example, the estimates presented for the September 10-11, 2017, SEP event demonstrated that the location-averaged effective dose rate due to the SEP at a height of 12 km was approximately $3\,\mu Sv/h$ (Kataoka et al., 2018). While this is less than the contribution due to galactic cosmic rays (more than twice that amount), it becomes comparable by an order of magnitude.

[1] `https://dmlab.cs.gsu.edu/rdp/ml-dataset.html`

Routine measurements of the radiation environment have become possible in the last decade via efforts from the Automated Radiation Measurements for Aerospace Safety experiment (ARMAS; Tobiska et al., 2015, 2016, 2018). The ARMAS device measures the local radiation environment conditions in real time during commercial aircraft flights. Currently, ARMAS has flown on more than 1000 flights and provided ∼400,000 individual measurements, sampling the aviation altitudes across the United States and worldwide[2]. The development and release of the Radiation Data Portal (RDP; Sadykov et al., 2021) provided a convenient overview and access to the ARMAS data.

Nowcasting or predicting the radiation environment in Earth's atmosphere can be carried out by physics-based approaches such as CARI (i.e., the Civil Aviation Research Institute; Copeland, 2021), Professional Aviation Dose Calculator (PANDOCA; Matthiä et al., 2014), the Nowcast of Aerospace Ionizing RAdiation System (NAIRAS; Mertens et al., 2013), etc. The NAIRAS model takes into account the galactic cosmic ray and SEP inputs. It is closely integrated with the ARMAS experiment by providing the assessments of the radiation dose rates for every ARMAS measurement. Recently, version 3 of the NAIRAS model was released (Mertens et al., 2023; Mertens et al., 2024) that integrates the barometric altitude corrections for a better comparison to ARMAS measurements at the appropriate altitudes, as well as run-on-request capabilities at the Community Coordinated Modeling Center (CCMC) by the National Aeronautics and Space Administration (NASA). The available amount of the ARMAS measurements makes it possible to explore data-driven, statistical, and machine learning (ML) based analysis for prediction conditions of the radiation environment. A comparison of data-driven and physics-based predictions is valuable for further improvements to the physics considered in the models.

Data preparation is the first, yet most demanding step when developing an ML model. Typically, datasets that can be easily fed into ML models with relatively low or no preparation required are called ML-ready datasets. Several works (Masson et al., 2024; Nita et al., 2022) have introduced the key principles required for these datasets, often related to the overall completeness, preparation levels, and accessibility to a wider community. The preparation of the ML-ready datasets suitable for radiation environment prediction could enhance community efforts in modeling the radiation environment and stimulate the development of data-driven approaches.

The primary goal of this paper is to present an ML-ready dataset that could be used for nowcasting and forecasting of the radiation environment at aviation altitudes. The paper is structured as follows: Section 2 describes the data mining and preprocessing steps required for both ARMAS and the space environment measurements. Section 3 presents steps for shaping the data into an ML-ready format. This includes the association of ARMAS measurements with the Geospace environment properties (such as ground-based neutron monitor counts, soft X-ray and proton flux measurements at the geostationary orbit, solar wind properties at L1 point, geomagnetic and global solar activity indexes) and the partitioning of multi-dimensional sparse data into three subsets that should be used directly for training and validation of the model. Section 4 provides a case example of how the dataset could be used for nowcasting, followed by a summary and discussion in Section 5.

2 Data Preparation

2.1 ARMAS data processing

The Automated Radiation Measurements for Aerospace Safety (ARMAS) experiment provides the richest publicly available dataset of atmospheric radiation measure-

[2] `https://dmlab.cs.gsu.edu/rdp/`

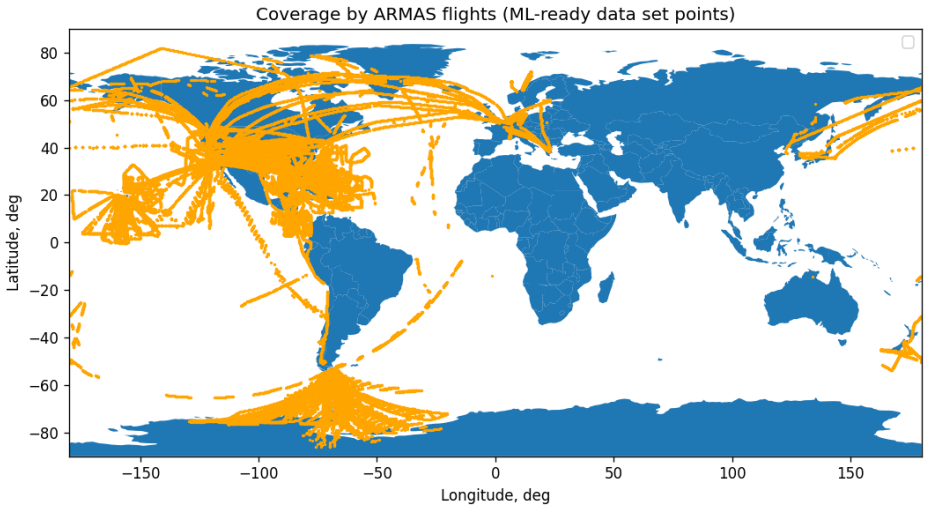

Figure 1. Coverage of considered ARMAS flight measurements over the Earth globe. The individual measurements are marked as orange points on the map.

ments along routes at aviation altitudes. In this work, we consider ARMAS data obtained from 1142 flight files (589 of which passed the selection criteria and contributed to an ML-ready dataset) from June 2013 to December 2023. This period of time covers the peak and decay of solar cycle 24, the solar minimum, and the rise phase of solar cycle 25, which enables sampling different levels of global solar activity. The flights cover a wide geographic span, including the extensive sampling of the continental US and territories, the Pacific Ocean regions, the North Atlantic, and Antarctica (Figure 1). ARMAS measurements are stored in the individual files for each flight and device and are accessible either via the Space Environment Technologies (SET) public data archive[3] or the Radiation Data Portal[4]. More details on ARMAS measurements can be found in Tobiska et al. (2015, 2016, 2018).

For each flight, we apply the following pre-processing steps. We sample the ARMAS measurements only in the barometric altitude range of 8.0 km – 15.5 km, as ARMAS often measures zero radiation below these altitudes. The corresponding maximum GPS altitude in the filtered dataset is 16.272 km. We notice that several flights had two or more ARMAS devices onboard, and the measurements of these devices were reported in separate files. For groups of files corresponding to the same flight, we consider only one file where ARMAS measurements are most strongly correlated with the NAIRAS v3 model estimates (in terms of Pearson correlation coefficient). We also remove the data points marked as non-science-use or during the periods of electromagnetic interference. The remaining data points are inspected manually for every flight. The procedure results in 589 unique flights propagating into the final dataset, and 92,476 unique data points.

The ARMAS files prepared by the SET public data archive[5] contain several additional parameters that we propagate into the final dataset. For example, there is information about the geomagnetic cutoff rigidities at the point of the flight measurement, as well as the barometric and GPS altitudes. The files also contain the physics-based nowcast of the radiation dose rate obtained with 2nd and 3rd versions of the Nowcast of At-

[3] `https://sol.spacenvironment.net/ARMAS/Archive/`
[4] `https://data.nas.nasa.gov/helio/portals/rdp/`
[5] `https://sol.spacenvironment.net/ARMAS_Archive/`

mospheric Ionizing Radiation for Aviation Safety model (NAIRAS; Mertens et al., 2013, 2023; Mertens et al., 2024). Version 3 (v3) of the model provides a better agreement with the ARMAS measurements and is therefore suggested to be used as a physics baseline for the radiation forecasts.

2.2 Neutron monitor stations

Along with the muons, secondary cosmic ray neutrons are an abundant remnant of the atmospheric particle cascades formed by the primary cosmic rays and, therefore, represent one of the most reliable ways to monitor the Geospace cosmic ray environment. The Neutron Monitor Database (NMDB, Kozlov et al., 2003; Väisänen et al., 2021) agglomerates the measurements from more than 50 neutron monitors around the globe. The NMDB Event Search Tool[6] (NEST) provides access to the data from these stations, including the count rates corrected with respect to the effects of the local atmospheric pressure.

The NAIRAS physics-based model utilizes neutron monitor data from four locations, Oulu (`OULU`, R_c=0.81 GeV), Lomnicky (`LKMS`, R_c=3.84 GeV), Thule (`THUL`, R_c=0.30 GeV), and Izmiran (`MOSC`, R_c = 2.43 GeV), as inputs. However, according to the NEST interface, `LKMS` and `MOSC` stations have some interruptions for the period of interest (2013-2023). The continuity of data is critical for the construction of the ML-ready dataset. Therefore, we decided to replace these stations with the neutron monitor station in Newark (`NEWK`, R_c=2.40 GeV), and we also added the South Pole station (`SOPO`) with extremely low geomagnetic cut-off rigidity (R_c=0.10 GeV). The pressure- and efficiency-corrected counts are obtained for all stations with a cadence of 1 minute.

2.3 Solar wind parameters

According to Tobiska et al. (2018), ARMAS measurements tend to have higher radiation dose rates than the predictions by the NAIRAS physics-based models, specifically in the regions of low geomagnetic cutoff rigidity. The presumed reason for this is the leakage of the energetic charged particles from the radiation belts, for which the regions of low cutoff rigidities are more accessible. Given that the perturbations of the Earth's magnetosphere and the dynamics of the radiation belts are tightly coupled with the solar wind parameters, we include the properties of the solar wind in our dataset.

The curated solar wind parameters are accessible via the OMNIWeb[7] online dataset with a 5-minute cadence. We include all major properties, such as the solar wind density, temperature, the magnitude, and all three components of the velocities and magnetic fields. All data from OMNIWeb is checked for continuity, and no affecting data gaps were found.

2.4 Energetic particles

Radiation levels at aviation altitudes might significantly increase (even exceed the galactic cosmic ray component) during solar energetic particle events (Kataoka et al., 2018). Tobiska et al. (2018) also noted that there are many enhanced radiation events in ARMAS data obtained for L-shells between 1.5 and 5 in the Western hemisphere, and indicated the radiation belt particles as a possible reason for this. Therefore, we incorporate the energetic proton and electron measurements in our database, all measured by the Geostationary Operational Environmental Satellite (GOES) series. The proton fluxes are obtained in 7 integrated energy channels: ≥1 MeV, ≥5 MeV, ≥10 MeV, ≥30 MeV,

[6] `https://www.nmdb.eu/nest/`
[7] `https://omniweb.gsfc.nasa.gov/`

≥50 MeV, ≥60 MeV, and ≥100 MeV. The data are obtained from the Integrated Space Weather Analysis (ISWA) webapp[8]. The energetic electron flux measurements are included only in one energy channel, ≥20 keV, because of the data continuity issues found in other energy channels. The resolution of both the proton and electron particle data is 5 min.

2.5 Solar X-ray measurements

Solar soft X-ray emission integrated over the solar disk is traditionally measured by the GOES spacecraft series, using the onboard X-ray sensor. We are utilizing 1-minute averaged fluxes in two channels, 1 – 8 Å and 0.5 – 4 Å, as input to our dataset. A similar type/cadence of the data was incorporated in the currently available Radiation Data Portal (Sadykov et al., 2021).

2.6 Geomagnetic activity indexes

We adopt hourly planetary Kp, Ap, and Dst indexes, all available via the OMNIWeb online dataset. The presented indices reflect the state of the Earth's magnetosphere during the major interplanetary disturbances, such as the presence of the Interplanetary Coronal Mass Ejections (ICMEs) or high-speed stream interaction regions in the solar wind. The variations in these parameters can affect the geomagnetic cutoff rigidity (Ptitsyna et al., 2021) and, therefore, modulate the galactic cosmic ray precipitation into the Earth's atmosphere.

2.7 Global solar activity and characteristics

In addition to the transient activity effects of the Sun on the Geospace, there are longer-term effects associated with the global solar activity. Although the neutron monitor data can capture effects related to the global solar activity (Section 2.2), the solar activity indexes make a picture of the evolution of the radiation environment more complete. For example, Koldobskiy et al. (2022) found the delay of about 7 months of the sunspot numbers with respect to the neutron monitor measurements. Therefore, we include the following global solar activity indicators:

- **Solar F10.7 index.** The Solar F10.7 index, which measures solar radio flux at a wavelength of 10.7 cm (2800 MHz), is a critical indicator of solar activity and has significant implications for Earth's atmospheric conditions.
- **Daily sunspot number**, which tracks the daily count of visible sunspots on the Sun's surface, is a traditional characteristic of solar activity level. Managed by the World Data Center SILSO[9] (Sunspot Index and Long-term Solar Observations), hosted at the Royal Observatory of Belgium, these data provide critical data for understanding solar cycles and their impacts on space weather.
- **Solar polar fields.** The solar polar magnetic fields, monitored by the Wilcox Solar Observatory[10] (WSO) at Stanford University, are crucial indicators of the Sun's magnetic activity and the solar cycle's progression. These fields, located at the Sun's poles, reverse approximately every 11 years, marking a new solar cycle. The strength and structure of the polar fields are key to predicting the magnitude of future solar cycles, as they play a central role in the generation of the Sun's magnetic field through the solar dynamo process. The WSO has provided near-continuous, detailed measurements of solar polar magnetic fields since 1976. It is an invalu-

[8] https://ccmc.gsfc.nasa.gov/tools/ISWA/
[9] https://www.sidc.be/SILSO/datafiles
[10] http://wso.stanford.edu/Polar.html

able resource for understanding solar dynamics, space weather prediction, and their influence on the heliosphere.

The illustration of some of these sources is presented in Figure 2. As one can see, most features are continuous across the time interval of interest. The only noticeable change occurs in high-energy proton flux ($\geq$50 MeV) with a jump in the background. However, since the protons of only a very high energy and flux could impact the radiation levels, the jumps in the proton flux background levels are not a point of concern.

3 ML-Ready Data Set Construction

Although ARMAS data and the corresponding Geospace environment characteristics have been collected and cleaned for the time interval of interest (June 2013 – December 2023), the data is still not ready for ML purposes. The ML-ready dataset must meet the following characteristics, such as data integrability and understandability. Specifically, the dataset should contain all components necessary for training and evaluating the ML models and results interpretability (data integrability) and should be accessible to non-domain experts with a reasonably low effort (data understandability). To meet the data integrability requirements, at least two more steps are required: the ARMAS measurements should be merged with the preceding environmental parameters, and the dataset has to be partitioned. To support the understandability of the data, partitions should be created with input from domain experts to prevent artificial correlations.

We associate individual ARMAS measurements with the temporally closest preceding measurements of the Geospace environment. We avoid using any temporal interpolations to avoid breaking the causality principle and ensure that we are not utilizing information from times in the future for the radiation nowcasting (Figure 3, left). Given that some parameters, like the daily sunspot numbers, have a relatively large time cadence (24 hours in this case), the preceding measurement of the daily sunspot number may represent the conditions up to 24 hours before the actual measurements by ARMAS.

The second step is the separation of the individual ARMAS measurements into the partitions (Figure 3, right). This process indicates the chunks of data that can be used for training, validation, and testing. This ensures the reproducibility of the research and the direct comparison between different models if trained on the same partitions. The partitioning previously was used in other ML-ready datasets, such as the Space Weather Analytics for Solar Flares, SWAN-SF (Angryk et al., 2020). The key point is that the ARMAS measurements cannot be separated randomly. The data points from the same flight could be just one minute away from each other. Therefore, the environmental properties are similar. Thus, if these data are split between the train and test partitions, this could generate artificial correlation between the partitions. This effect was previously recognized as a 'temporal coherence' for the problem of flare forecasting (Ahmadzadeh et al., 2021). Therefore, the data from every ARMAS flight must be allocated to the same partition.

In principle, each partition should represent the entire dataset, meaning the distribution of the parameters in each partition should represent the entire dataset (Liu et al., 2019). Separation into partitions could be a challenging problem if data coverage is sparse, and the problem is multi-dimensional. Unfortunately, this is the case for the considered dataset: the measurements are acquired along slightly more than 1000 individual flight trajectories, and more than 40 Geospace parameters are associated with every measurement. To overcome the manual trial-and-error partitioning process, we utilized an ML clustering algorithm to help with it. The steps are as follows:

1. **Down-selection of the five parameters for clustering.** As was indicated above, the number of Geospace parameters is relatively large. Distance-based clustering

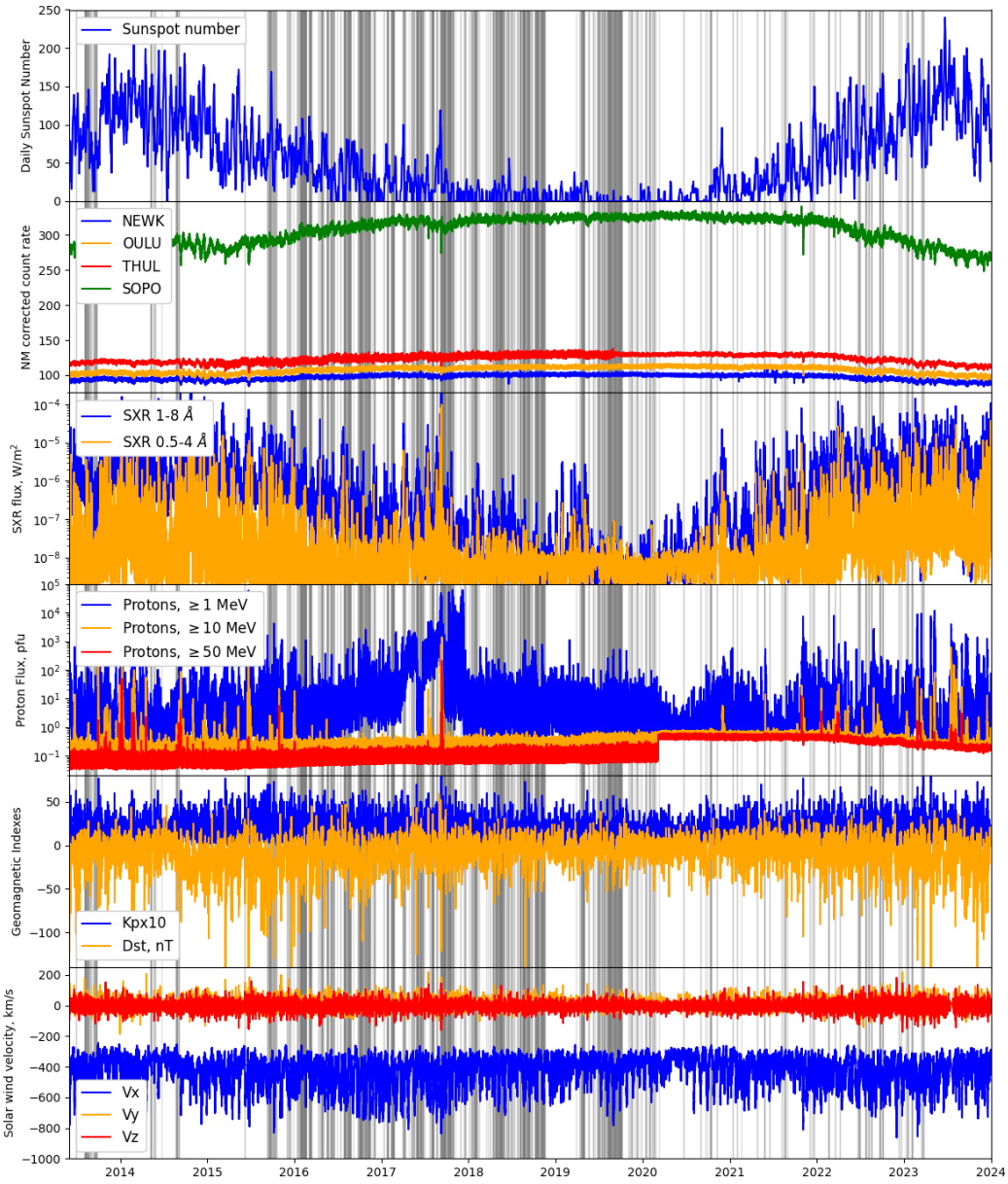

Figure 2. Evolution of some selected Geospace environment parameters from June 2013 until December 2024, from top to bottom: daily sunspot number, neutron monitor corrected counts from four stations considered, integrated soft X-ray fluxes, energetic proton fluxes, geomagnetic indexes (Kp and Dst), and solar wind velocities at L1. Gray lines in the background represent the time moments covered by the ARMAS flight measurements.

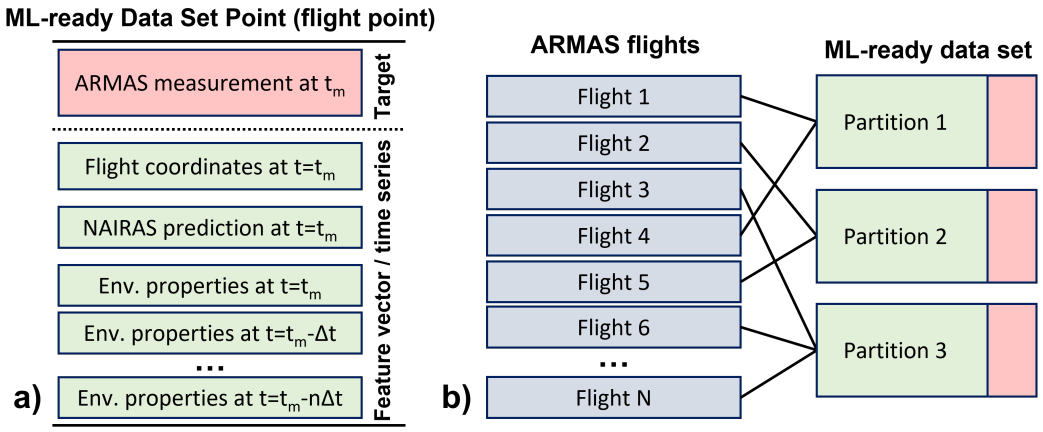

Figure 3. (a) Schematic structure of the ML-ready dataset entity. A target corresponds to the measurement of radiation dose rate during the ARMAS flight, and the feature vector represents the flight timing and coordinates, NAIRAS predictions, and prehistory of the measurements of the environment. (b) Illustration of the subdivision of ARMAS flights into partitions.

algorithms struggle with high dimensionality (the so-called 'curse of dimensionality' problem). However, many of the parameters are expected to be correlated with each other. For example, there is a strong correlation between all geomagnetic indices (such as Kp, Ap, Dst), global solar activity parameters (daily sunspot numbers, F10.7 flux, polar magnetic field measurements), description of the location in geomagnetic coordinates (geomagnetic latitude, L-shell, cutoff rigidity), etc. Instead of considering all of them, we down-select five parameters based on which we will cluster the data. The first three are related to the location of the measurement: geomagnetic longitude, geomagnetic latitude, and barometric flight altitude. The fourth parameter describes the geomagnetic activity, for which we use the Dst index. The last parameter we use is the daily sunspot number that reflects global solar activity levels. Although this selection is not unique, it is sufficient for partitioning purposes. Because the number of ARMAS flights during the solar energetic particle (SEP) events is very low, we did not use the SEP-related measurements for partitioning.

2. **Clustering of individual measurements.** We apply the Gaussian Mixture Model (GMM) clustering based on the five parameters for the individual ARMAS measurements. The number of GMM components was selected to be 100, which is sufficiently large to create a relatively equal representation for all parameters in every partition. We note here that the GMM is a soft clustering methodology as it assigns the probabilities of every point to belong to a certain cluster rather than associating it with a particular cluster.
3. **Associating flights with clusters.** Each individual measurement in each flight now has a probability (or a weight) of belonging to each of the 100 clusters. To associate the entire flight with the particular cluster, we sum up the probabilities of the individual measurements in this flight for every cluster. Therefore, a flight becomes assigned entirely to a particular cluster for which the sum of the probabilities of the individual measurements is the highest. We perform these assignments for all flights in our dataset.
4. **Separation of the flights into three partitions.** We start from the first cluster and distribute the flights to this cluster into three partitions in sequential order. After distributing all flights in the cluster, we move to the next cluster and repeat the procedure starting from the next partition. For example, if cluster 1

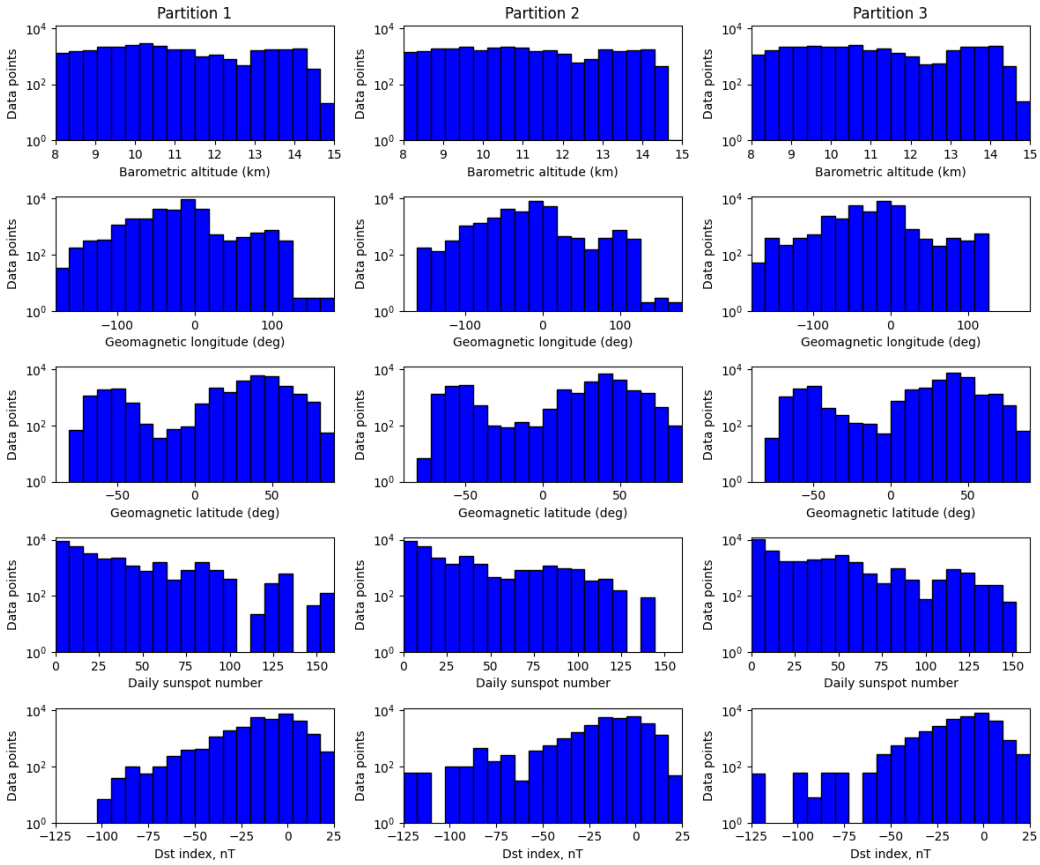

Figure 4. Distribution of the parameters used for clustering the data points (barometric altitude, geomagnetic longitude, geomagnetic latitude, daily sunspot number, and Dst index) within each partition of the dataset. Each row corresponds to a single parameter. The partition is indicated in the header of each column.

has four flights, then flight #1 will be distributed to partition 1, flight #2 – to partition 2, flight #3 – to partition 3, and flight #4 – again to partition 1. Then one moves to cluster 2, and flight #1 in this cluster goes now to partition 2, flight #2 – to partition 3, etc. This procedure ensures that all partitions will include flights belonging to the same cluster (and, therefore, likely sampling similar spatial locations and geomagnetic and solar activity levels).

While the partitioning strategy does not necessarily result in the 'most optimal' distribution of the flights, it leads to a satisfactory representation of parameter combinations in every partition while avoiding the brute-force partitioning. The result of the distribution of the parameters within each partition is presented in Figure 4. The histograms of the parameters are relatively similar for every column of the partitioning, indicating that each partition samples the parameter space relatively equally. Some differences are evident only for the extreme values of the parameters, such as the Dst indexes at around $\sim$ -100 nT or the daily sunspot numbers of $\sim$150, which occur because not too many ARMAS flights have occurred during such high activity levels.

It is important to check if other parameters are sampled evenly across the partitions. Figure 5 illustrates the distributions of the parameters in three partitions that have

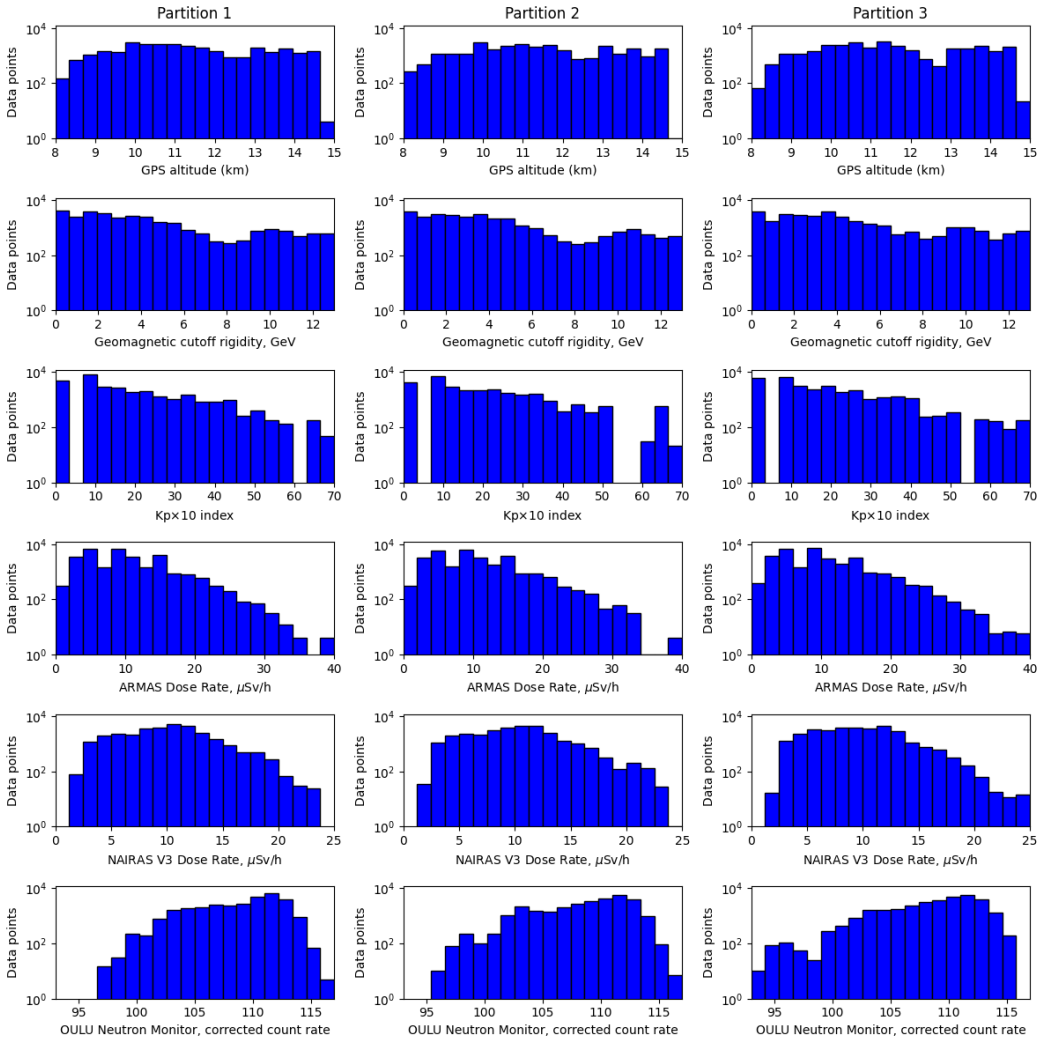

Figure 5. Distribution of the parameters that were not used for the clustering of the data points (GPS altitude, geomagnetic cutoff rigidity, Kp index multiplied by 10, ARMAS dose rate, NAIRAS dose rate, and corrected count rates of secondary cosmic ray neutrons detected by `OULU` station) within each partition of the dataset. Each row corresponds to a single parameter. The partition is indicated in the header of each column.

not directly participated in the clustering. The distribution of parameters is very similar, indicating a successful partitioning process.

The resulting data represents the ML-ready dataset that utilizes the last 'snapshot' of Geospace properties before the ARMAS measurement. The ML-ready dataset consists of a feature vector (a set of characteristics based on which the prediction is made, representing Geospace parameters) and a target (a characteristic to predict, here ARMAS measurement). The feature vector comprises the flight timing and location properties, as well as the most recent properties of the environment. In addition to the most recent properties (Figure 3), one can provide the time series of the evolution of the Geospace parameters up to the time before the measurements. The preceding properties of the environment can be forward-interpolated with the cadence Δt for n time steps before the ARMAS measurement. At the same time, the partitioning of the individual ARMAS mea-

surements and related time series remains as described above. Following this, we have finally constructed three publicly available ML-ready datasets:

- The dataset that represents the most recent Geospace measurements and does not involve their time series ('static' dataset, $n = 0$);
- The dataset that includes a 1-hour prehistory of the Geospace measurements before ARMAS flight measurement ('dynamic' dataset 1, $n = 12$ and $\Delta t = 5\,\text{min}$);
- The dataset that includes a 24-hour prehistory of the Geospace measurements before ARMAS flight measurement ('dynamic' dataset 2, $n = 24$ and $\Delta t = 1\,\text{hour}$);

All three versions of the datasets are currently accessible via the Radiation Data Portal[11]

4 Data Set Use Case Example

In this section, we illustrate how the constructed datasets can be utilized for ML-driven forecasting of atmospheric radiation. For this demonstration, we proceed with the simplest version of the three datasets constructed and described above, the 'static' ML-ready dataset, which includes only the latest point-in-time measurement of every Geospace parameter. We use partition 1 of this dataset to train the ML model described below, and partition 2 for the evaluation of the performance of the model and its comparison with the predictions of the NAIRAS-v3 physics-based model. Among the ML approaches available off-the-shelf, we select the Random Forest regressor (Breiman, 2001), a bagging tree-based ensemble learning algorithm. Random Forest has previously been successfully applied to a variety of classification and regression problems in the space physics domain, including the prediction of solar flares and solar energetic particle events (Liu et al., 2017; O'Keefe et al., 2022), forecasting the duration of enhanced soft X-ray radiation during the flare (Reep & Barnes, 2021), the timing of the solar wind propagation (Baumann & McCloskey, 2021), ion-kinetic instability detection in the solar wind-like plasmas (Sadykov et al., 2025), etc. We note here that, despite the promising track record, one cannot guarantee that the Random Forest approach is the most optimal for the considered problem. Therefore, the survey of other ML models is highly encouraged and is one of the authors' goals for the future.

We have used the Random Forest model available at the `scikit-learn` Python library package (Pedregosa et al., 2011). The model has several hyperparameters to optimize, such as the number of individual decision trees in the ensemble, the maximum depth restriction for each tree, the number of features randomly selected and propagated into every tree, etc. Typically, the hyperparameters are fine-tuned during the cross-validation phase when the models of different hyperparameters are evaluated on a designated partition or a subset of the training partition. Here, we do not perform the detailed cross-validation but rather use the parameters we found to perform satisfactorily for the considered study during our preliminary tests. Our Random Forest consists of 100 decision trees of a depth of 10 or less, with at least 2 samples from the dataset required for the split within the tree, and at least 4 samples to be at the leaf node. To guide the training process, we utilize the mean squared error as a measure of the regressor's performance.

The results are presented in Figure 6. The left panel illustrates the scatterplot of the radiation dose rates predicted by the ML-driven model against the ARMAS measurements used as a ground truth. One can see that points tend to be mostly organized along the red line, which represents the ideal one-to-one prediction. It is also visible that the predictions demonstrate a stronger spread and systematic deviation from the ideal prediction line for the larger values of dose rates. In fact, the part of the distribution cov-

[11] `https://dmlab.cs.gsu.edu/rdp/ml-dataset.html`

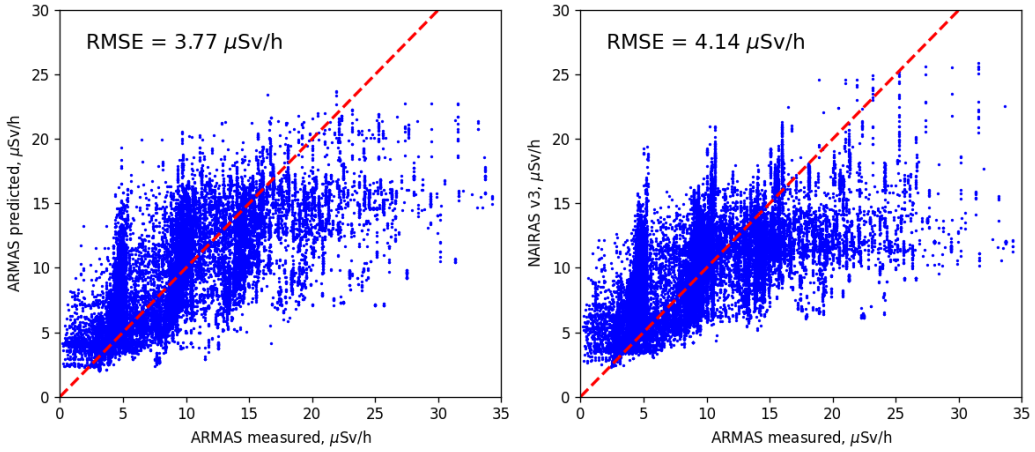

Figure 6. Left: Measured radiation dose rates VS predicted using an ML model (Random Forest Regressor). The ML model was trained on partition 1 of the static dataset and evaluated on partition 2. Right: Measured radiation dose rates VS nowcast of a physics-based NAIRAS V3 model for partition 2.

ering the dose rates of $>15\,\mu Sv/h$ is mostly situated below the perfect prediction line, indicating that the ML-driven predictions typically underestimate the actual dose rates in the cases where the measured radiation doses are high. Interestingly, the comparison of ARMAS measurements with the nowcast of the physics-based NAIRAS-v3 model (presented in the right panel of Figure 6) demonstrates the same pattern. The root mean squared error (RMSE) computed for the ML-driven prediction is $3.77\,\mu Sv/h$, whereas it is $4.14\,\mu Sv/h$ for the NAIRAS-v3 model. One can see that the ML-driven model seems to deliver slightly more accurate predictions both in terms of the RMSE measure across the test partition and based on the qualitative appearance of the scatterplots in Figure 6. Overall, this emphasizes the potential for ML approaches in nowcasting the radiation at aviation altitudes. The potential next steps can include generalizing the results over various train-validation-test partition combinations, considering other ML algorithms, and involving the time series data of Geospace parameters for forecast improvement. The constructed ML-ready datasets enable such investigations, opening the perspective of the new studies of the research community.

5 Summary and Discussion

This paper presents the construction of the ML-ready dataset for nowcasting atmospheric radiation at aviation altitudes. We have leveraged the publicly available effective dose rate measurements by the ARMAS device acquired over 589 flights, and created a pre-partitioned ML-ready dataset, which requires minimal preprocessing to be used for ML purposes. Some of the dataset features are summarized below:

- The resulting dataset comprises 589 ARMAS flights containing 92,476 effective dose rate measurements. While the flights are mostly accomplished on top of the continental US and territories (Figure 1), the dataset also samples the regions over the Pacific Ocean, the North Atlantic region, and Antarctica.
- The radiation measurements are supported by the measurements from four neutron monitor stations (`OULU`, `DOMC`, `NEWK`, and `THUL`), solar wind properties at L1, measurements of proton and soft X-ray fluxes by GOES spacecraft series, geomag-

netic activity indexes, and global solar activity characteristics (such as sunspot numbers, F10.7 flux, and solar polar fields). This provides an opportunity for comprehensive studies of the dependencies of the radiation environment on the Geospace drivers.

- The dataset has been pre-partitioned into three subsets, which can be directly used for training, validation, and testing purposes. During pre-partitioning, it was ensured that the data points from the same flight are within the same partition, and that the sampling of the parameter space is more or less the same within any partition (see Figures 4 and 5 for parameter distributions).
- Three versions of the dataset are constructed. The 'static' version includes the most recent properties of the environment only. The 'dynamic' versions include the prehistory of Geospace parameters as time series. The 1-hour and 24-hour long prehistory is considered.
- The use case example demonstrates on the selected test subset the possibility of constructing an ML model that predicts the effective dose rates with the average root mean squared error of 3.77μSv/h, which is slightly better than the NAIRAS v3 physics-based model nowcast (4.14μSv/h). This holds promise for the development of ML-driven models of radiation forecasting in the future.

We envision that the constructed datasets could be used for various scientific applications. First, the problem of the atmospheric radiation nowcast given the state of the environment requires a detailed evaluation with respect to the ML algorithms and involvement of the time series. As highlighted in Section 4, even the static version of the dataset with the Random Forest regressor demonstrates the promising results. The nowcasting and forecasting results can vary with the algorithm (Ali et al., 2024; Goodwin et al., 2024; O'Keefe et al., 2024); therefore, the consideration of other ML methods is necessary. The benefits of including time series properties need to be assessed as well. Second, the slight manipulation of the dataset allows us to consider a forecasting problem instead of nowcasting. All one has to do is consider the 'dynamic' dataset versions and avoid considering the data within a certain latency window before the ARMAS measurement. Obviously, the length of the time series (1 h and 24 h) would limit the considered latency windows. Third, the dataset could enhance the understanding of radiation environment physics. Besides the standard correlation analyses possible, one could investigate feature importance to understand the influence of some Geospace parameters on the radiation environment (e.g., Yeolekar et al., 2021; Sadykov & Kosovichev, 2017). The developed dataset opens a promising number of prospective studies and facilitates the development of models for the aviation radiation domain.

Open Research

The developed machine learning-ready dataset to nowcast the effective dose rates at aviation altitudes is currently publicly available via the Radiation Data Portal (`https://dmlab.cs.gsu.edu/rdp/ml-dataset.html`). The original ARMAS data files are publicly available from the ARMAS Data Archive at Space Environment Technologies (`https://sol.spacenvironment.net/ARMAS/Archive/`). The Neutron Monitor can be accessed via the NMDB database (`https://www.nmdb.eu/`). The solar wind measurements, geomagnetic activity indexes, and F10.7 index data are publicly accessible via the OMNIWeb service (`https://omniweb.gsfc.nasa.gov/`). The GOES energetic particle flux and soft X-ray data can be accessed via the National Oceanic and Atmospheric Administration National Centers for Environmental Information Archive (NOAA NCEI, `https://www.ncei.noaa.gov/`). We thank the developers of the Integrated Space Weather Analysis (ISWA, `https://ccmc.gsfc.nasa.gov/tools/ISWA/`) systems API for the possibility of retrieving the GOES integrated proton fluxes since 2020. The daily sunspot number is obtained via the publicly accessible World Data Center SILSO (`https://www.sidc`

.be/SILSO/datafiles). The solar polar field measurements are publicly available via the Wilcox Solar Observatory website (http://wso.stanford.edu/Polar.html).

Acknowledgments

This work has been supported by the NASA HITS grant 80NSSC22K1561. VMS also thanks the NSF FDSS grant 1936361 and NASA LWS grant 80NSSC24K1111. We acknowledge the NMDB database (www.nmdb.eu), founded under the European Union's FP7 programme (contract no. 213007) for providing data, and the PIs of individual neutron monitors at: Newark and Thule (University of Delaware Department of Physics and Astronomy and the Bartol Research Institute, USA), Kerguelen (Observatoire de Paris and the French Polar Institute IPEV, France), Oulu (Sodankyla Geophysical Observatory of the University of Oulu, Finland), South Pole (University of Wisconsin, River Falls, USA)

References

Ahmadzadeh, A., Aydin, B., Georgoulis, M. K., Kempton, D. J., Mahajan, S. S., & Angryk, R. A. (2021, June). How to Train Your Flare Prediction Model: Revisiting Robust Sampling of Rare Events. *The Astrophysical Journal Supplement Series*, *254*(2), 23. doi: 10.3847/1538-4365/abec88

Ali, A., Sadykov, V., Kosovichev, A., Kitiashvili, I. N., Oria, V., Nita, G. M., ... Marroquin, R. D. (2024, January). Predicting Solar Proton Events of Solar Cycles 22-24 Using GOES Proton and Soft-X-Ray Flux Features. *The Astrophysical Journal Supplement Series*, *270*(1), 15. doi: 10.3847/1538-4365/ad0a6c

Angryk, R. A., Martens, P. C., Aydin, B., Kempton, D., Mahajan, S. S., Basodi, S., ... Georgoulis, M. K. (2020, January). Multivariate time series dataset for space weather data analytics. *Scientific Data*, *7*(1), 227. doi: 10.1038/s41597-020-0548-x

Baumann, C., & McCloskey, A. E. (2021, June). Timing of the solar wind propagation delay between L1 and Earth based on machine learning. *Journal of Space Weather and Space Climate*, *11*, 41. doi: 10.1051/swsc/2021026

Breiman, L. (2001, January). Random Forests. *Machine Learning*, *45*, 5-32. doi: 10.1023/A:1010933404324

Cho, G., Kim, J. H., Park, T. S., & Cho, K. (2017). Proposing a simple radiation scale for the public: radiation index. *Nuclear Engineering and Technology*, *49*(3), 598–608.

Copeland, K. (2021, March). *Cari-7 documentation: Radiation transport in the atmosphere* (Tech Report No. DOT/FAA/AM-21/05). United States. Department of Transportation. Federal Aviation Administration. Office of Aviation. Civil Aerospace Medical Institute. Retrieved from https://rosap.ntl.bts.gov/view/dot/57224

Goodwin, G. T., Sadykov, V. M., & Martens, P. C. (2024, April). Investigating Performance Trends of Simulated Real-time Solar Flare Predictions: The Impacts of Training Windows, Data Volumes, and the Solar Cycle. *The Astrophysical Journal*, *964*(2), 163. doi: 10.3847/1538-4357/ad276c

Kataoka, R., Sato, T., Miyake, S., Shiota, D., & Kubo, Y. (2018, July). Radiation Dose Nowcast for the Ground Level Enhancement on 10-11 September 2017. *Space Weather*, *16*(7), 917-923. doi: 10.1029/2018SW001874

Koldobskiy, S. A., Kähkönen, R., Hofer, B., Krivova, N. A., Kovaltsov, G. A., & Usoskin, I. G. (2022, March). Time Lag Between Cosmic-Ray and Solar Variability: Sunspot Numbers and Open Solar Magnetic Flux. *Solar Physics*, *297*(3), 38. doi: 10.1007/s11207-022-01970-1

Kozlov, V., Kudela, K., Starodubtsev, S., Turpanov, A., Usoskin, I., & Yanke, V.

(2003, September). Neutron monitor database in real time. In A. Wilson (Ed.), *Solar variability as an input to the earth's environment* (Vol. 535, p. 675-678).

Liu, C., Deng, N., Wang, J. T. L., & Wang, H. (2017, July). Predicting Solar Flares Using SDO/HMI Vector Magnetic Data Products and the Random Forest Algorithm. *The Astrophysical Journal*, *843*(2), 104. doi: 10.3847/1538-4357/aa789b

Liu, H., Chen, S.-M., & Cocea, M. (2019). Subclass-based semi-random data partitioning for improving sample representativeness. *Information Sciences*, *478*, 208-221. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0020025518308910` doi: https://doi.org/10.1016/j.ins.2018.11.002

Masson, A., Fung, S. F., Camporeale, E., Kuznetsova, M. M., Poedts, S., Barnum, J., ... Cecconi, B. (2024). Heliophysics and space weather information architecture and innovative solutions: Current status and ways forward. *Advances in Space Research*. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0273117724004939` doi: https://doi.org/10.1016/j.asr.2024.05.052

Matthiä, D., Meier, M. M., & Reitz, G. (2014, March). Numerical calculation of the radiation exposure from galactic cosmic rays at aviation altitudes with the PANDOCA core model. *Space Weather*, *12*(3), 161-171. doi: 10.1002/2013SW001022

Mertens, C. J., Gronoff, G. P., Zheng, Y., Buhler, J., Willis, E., Petrenko, M., ... Minow, J. (2024). Nairas atmospheric and space radiation environment model. *IEEE Transactions on Nuclear Science*, *71*(4), 618-625. doi: 10.1109/TNS.2023.3330675

Mertens, C. J., Gronoff, G. P., Zheng, Y., Petrenko, M., Buhler, J., Phoenix, D., ... Minow, J. (2023, May). NAIRAS Model Run-On-Request Service at CCMC. *Space Weather*, *21*(5), e2023SW003473. doi: 10.1029/2023SW003473

Mertens, C. J., Meier, M. M., Brown, S., Norman, R. B., & Xu, X. (2013, October). NAIRAS aircraft radiation model development, dose climatology, and initial validation. *Space Weather*, *11*(10), 603-635. doi: 10.1002/swe.20100

Nita, G., Ahmadzadeh, A., Criscuoli, S., Davey, A., Gary, D., Georgoulis, M., ... Wang, J. T. L. (2022, March). Revisiting the Solar Research Cyberinfrastructure Needs: A White Paper of Findings and Recommendations. *arXiv e-prints*, arXiv:2203.09544. doi: 10.48550/arXiv.2203.09544

O'Keefe, P. M., Sadykov, V., Kosovichev, A., Kitiashvili, I. N., Oria, V., Nita, G. M., ... Marroquin, R. D. (2024, December). The Random Hivemind: An ensemble deep learning application to the solar energetic particle prediction problem. *Advances in Space Research*, *74*(12), 6252-6263. doi: 10.1016/j.asr.2024.04.044

O'Keefe, P. M., Sadykov, V. M., Kosovichev, A. G., Nita, G. M., Oria, V., Sharma, S., ... Jie Chong, C. (2022, June). *Handling Highly Imbalanced Data in Machine Learning Applications.* Zenodo. doi: 10.5281/zenodo.6780972

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ptitsyna, N. G., Danilova, O. A., Tyasto, M. I., & Sdobnov, V. E. (2021, March). Dynamics of Cosmic-Ray Cutoff Rigidity and Magnetospheric Parameters during Different Phases of the Storm of November 20, 2003. *Geomagnetism and Aeronomy*, *61*(2), 169-179. doi: 10.1134/S0016793221010114

Reames, D. V. (2021). *Solar Energetic Particles. A Modern Primer on Understanding Sources, Acceleration and Propagation* (Vol. 978). doi: 10.1007/978-3-030-66402-2

Reep, J. W., & Barnes, W. T. (2021, October). Forecasting the Remaining Duration of an Ongoing Solar Flare. *Space Weather*, *19*(10), e02754. doi: 10.1029/

2021SW002754
Sadykov, V. M., Kitiashvili, I. N., Tobiska, W. K., & Guhathakurta, M. (2021, January). Radiation Data Portal: Integration of Radiation Measurements at the Aviation Altitudes and Solar-Terrestrial Environment Observations. *Space Weather*, *19*(1), e2020SW002653. doi: 10.1029/2020SW002653
Sadykov, V. M., & Kosovichev, A. G. (2017, November). Relationships between Characteristics of the Line-of-sight Magnetic Field and Solar Flare Forecasts. *The Astrophysical Journal*, *849*(2), 148. doi: 10.3847/1538-4357/aa9119
Sadykov, V. M., Ofman, L., Boardsen, S. A., Yogesh, Mostafavi, P., Jian, L. K., ... Martinović, M. (2025, May). Identification of Ion-Kinetic Instabilities in Hybrid-PIC Simulations of Solar Wind Plasma with Machine Learning. *arXiv e-prints*, arXiv:2505.18271. doi: 10.48550/arXiv.2505.18271
Tobiska, W. K., Atwell, W., Beck, P., Benton, E., Copeland, K., Dyer, C., ... Xapsos, M. A. (2015, April). Advances in Atmospheric Radiation Measurements and Modeling Needed to Improve Air Safety. *Space Weather*, *13*(4), 202-210. doi: 10.1002/2015SW001169
Tobiska, W. K., Bouwer, D., Smart, D., Shea, M., Bailey, J., Didkovsky, L., ... Yoon, K. (2016, November). Global real-time dose measurements using the Automated Radiation Measurements for Aerospace Safety (ARMAS) system. *Space Weather*, *14*(11), 1053-1080. doi: 10.1002/2016SW001419
Tobiska, W. K., Didkovsky, L., Judge, K., Weiman, S., Bouwer, D., Bailey, J., ... Fuschino, R. (2018, October). Analytical Representations for Characterizing the Global Aviation Radiation Environment Based on Model and Measurement Databases. *Space Weather*, *16*(10), 1523-1538. doi: 10.1029/2018SW001843
Väisänen, P., Usoskin, I., & Mursula, K. (2021, May). Seven Decades of Neutron Monitors (1951-2019): Overview and Evaluation of Data Sources. *Journal of Geophysical Research (Space Physics)*, *126*(5), e28941. doi: 10.1029/2020JA02894110.1002/essoar.10505091.1
Yeolekar, A., Patel, S., Talla, S., Puthucode, K. R., Ahmadzadeh, A., Sadykov, V. M., & Angryk, R. A. (2021, December). Feature Selection on a Flare Forecasting Testbed: A Comparative Study of 24 Methods . In *2021 international conference on data mining workshops (icdmw)* (p. 1067-1076). Los Alamitos, CA, USA: IEEE Computer Society. Retrieved from `https://doi.ieeecomputersociety.org/10.1109/ICDMW53433.2021.00138` doi: 10.1109/ICDMW53433.2021.00138