Evaluating Proton Intensities for the SMILE Mission

Simon Patrick Mischel[1], Elena A. Kronberg[2], and Christopher Philippe Escoubet[3]

[1]Ludwig-Maximilians-Universität München
[2]Ludwig Maximilian University of Munich
[3]ESA / ESTEC

April 26, 2024

Abstract

This study introduces five linear regression models developed to accurately predict proton intensities in the critical energy range of 92.2 keV to 159.7 keV. To achieve this task we utilized 14 years of data sourced from the Cluster's RAPID experiment and NASA's OMNI database. This data was then aligned with the Solar wind-Magnetosphere-Ionosphere Link Explorer (SMILE) mission's trajectory, to increase model accuracy in the relevant regions. Our approach diverges from existing methodologies by offering a user-friendly model that doesn't require specialized software, making it accessible for broader applications in satellite mission planning and risk assessment. The research segregates the dataset into four distinct regions, each analyzed for proton intensity dynamics. In the outer regions (|YGSE| [?] 6 Re) there is a pronounced dependence on radial distance and solar wind speed. In contrast, the inner regions (|YGSE| [?] 6 Re) demonstrate a significant dependence of proton intensities on the absolute value of the z-coordinate and the magnetic field line topology. Our models achieved a Spearman correlation ranging from 0.57 to 0.72 on the test set, indicating good predictive capabilities. The findings emphasize the role of regional characteristics in space weather prediction and underscore the potential for tailored approaches in future research.

Evaluating Proton Intensities for the SMILE Mission

Simon Mischel[1], Elena A. Kronberg[1], C.P. Escoubet[2]

[1]Department of Earth and Environmental Sciences (Geophysics), Ludwig Maximilian University of Munich (LMU) Munich, Theresienstr. 41, Munich, D-80333, Germany
[2]European Space Research and Technology Centre, Noordwjik, Keplerlaan 1, 2201 AZ, The Netherlands

Key Points:

- Developed models predict proton intensities impacting satellites, aiding space weather forecasting and mission planning.
- Different regions in space showcase distinct relations between proton intensities and predicting parameters.
- Study findings highlight the importance of tailored approaches in space weather prediction.

Corresponding author: Simon Mischel, simonmischel@hotmail.com

Abstract

This study introduces five linear regression models developed to accurately predict proton intensities in the critical energy range of 92.2 keV to 159.7 keV. To achieve this task we utilized 14 years of data sourced from the Cluster's RAPID experiment and NASA's OMNI database. This data was then aligned with the Solar wind-Magnetosphere-Ionosphere Link Explorer (SMILE) mission's trajectory, to increase model accuracy in the relevant regions. Our approach diverges from existing methodologies by offering a user-friendly model that doesn't require specialized software, making it accessible for broader applications in satellite mission planning and risk assessment. The research segregates the dataset into four distinct regions, each analyzed for proton intensity dynamics. In the outer regions ($|\mathrm{YGSE}| \geq 6\,R_e$) there is a pronounced dependence on radial distance and solar wind speed. In contrast, the inner regions ($|\mathrm{YGSE}| \leq 6\,R_e$) demonstrate a significant dependence of proton intensities on the absolute value of the z-coordinate and the magnetic field line topology. Our models achieved a Spearman correlation ranging from 0.57 to 0.72 on the test set, indicating good predictive capabilities. The findings emphasize the role of regional characteristics in space weather prediction and underscore the potential for tailored approaches in future research.

Plain Language Summary

We developed a new model to predict space weather, specifically focusing on proton intensities, which can impact how well satellites work in space. We used 14 years of space observations to create five easy-to-use numerical models. These models are designed to help with planning and protecting future satellite missions, such as the upcoming SMILE mission, from space weather effects. In our study, we looked closely at different areas in space around Earth. We found that in the outer areas ($|\mathrm{YGSE}| \geq 6\,R_e$), the distance from Earth and the speed of the solar wind are important for understanding proton behavior. However, in areas ($|\mathrm{YGSE}| \leq 6\,R_e$), the height above Earth (measured along the z-direction) and the type of magnetic field lines play a more significant role. This shows us that different areas in space around Earth can be affected by space weather in different ways. Our models did a good job of predicting these effects, showing that choosing a tailored approach can be useful when forecasting proton intensities.

1 Introduction

Space weather events, driven by solar activities pose significant challenges to satellite operations and measurements. Notable examples include the European Space Agency's (ESA) Cluster and X-ray Multi-Mirror (XMM-Newton) missions. The Cluster mission, particularly its Research with Adaptive Particle Imaging Detector (RAPID)/Imaging Electron Spectrometer (IES), has encountered challenges due to high proton intensities leading to measurement contamination (Wilken et al., 1997; Kronberg et al., 2016; Kronberg, Daly, et al., 2021). Similarly, the X-ray telescope aboard the XMM-Newton spacecraft experienced significant operational disruptions, with approximately 40% of its observation time compromised due to background contamination (Walsh et al., 2014). Furthermore, investigations into the XMM-Newton telescope's susceptibility to soft protons highlight proton intensities in the sub-100 to 300 keV range, particularly around 100 keV, as the most damaging, leading to significant operational challenges and data contamination, see (Fioretti et al., 2016) and references therein.

The upcoming European-Chinese Solar wind-Magnetosphere-Ionosphere Link Explorer (SMILE) mission (Branduardi-Raymont et al., 2018), slated for launch in 2025, aspires to deepen our understanding of the Sun-Earth interaction, decoding space weather hazards and understanding energy entry into Earth's magnetosphere. While missions like ATHENA (Advanced Telescope for High Energy Astrophysics), which will maintain an orbit around the L2 Lagrange point with a continuously large distance to Earth, are ex-

pected to face minimal threats from soft protons (Perinati et al., 2024), SMILE with its highly inclined elliptical orbit around earth, akin to that of Cluster and XMM-Newton, will navigate through diverse magnetospheric regions, making it susceptible to soft proton radiation. Particularly its Soft X-ray Imager (SXI) telescope, presents challenges concerning radiation exposure (Raab et al., 2016; Branduardi-Raymont & Wang, 2022). Therefore a critical component of achieving SMILE's objectives is the accurate prediction of proton radiation levels, which significantly affect the Total Ionizing Dose (TID) and Total Non-Ionizing Dose (TNID) absorbed by the Charge-Coupled Devices (CCDs) of the Soft X-ray Imager (SXI) (Hubbard et al., 2024). In recent discussions (M. Hubbard, personal communication, 2023) the necessity for models that accurately estimate radiation levels was emphasized, particularly in critical energy ranges below 300 keV, crucial for the SMILE mission's success. A strong preference for models that are not only accurate but also straightforward and interpretable was expressed.

To meet these needs, our study adopts a distinct approach compared to existing research, which is based on machine learning black box models, such as the works of Kronberg et al. (2020) and Kronberg, Hannan, et al. (2021). We aim to develop a simple, user-friendly linear regression model leveraging data from the Cluster mission and NASA's OMNI database (King & Papitashvili, 2005). Our model's simplicity and ease of use make it accessible to a broader range of users, not requiring specialized software or extensive computational resources. This approach not only contributes to the scientific understanding of space weather phenomena but also offers practical tools for satellite mission planning and risk assessment.

2 Data Analysis and Processing

2.1 Data Preparation: Adapting CLUSTER's Dataset for the SMILE Mission's Trajectory

The proton intensity data for this research was taken from the Cluster's RAPID experiment, ranging from 2001 to 2015. The experiment captures 3-D energetic electron and ion fluxes above approximately 30 keV using the Imaging Electron Spectrometer (IES) and the Imaging Ion Mass Spectrometer (IIMS) instruments. Situated in the SCENIC detector head, the IIMS instrument identifies ion energies and species. The methodology involves using start and stop signals produced from electrons emitted by an initial thin foil on the solid-state detector's surface. The time-of-flight (TOF) between these signals, combined with the known energy, discerns the species and energy channel (Daly & Kronberg, 2023). This study specifically uses data from the *p3* channel, targeting proton intensities between 92.2 keV and 159.7 keV. As predicting parameters for solar, solar wind and geomagnetic activity we used variables from the OMNI database(`https://omniweb.gsfc.nasa.gov/`), see also King and Papitashvili (2005).

Aiming for a model tailored to the SMILE mission's trajectory (see Figure 1), data filtering was imperative. Points not adhering to the following spatial parameters were excluded: $-10.5R_e \leq x \leq 11.2R_e$; $-10.8R_e \leq y \leq 11.5R_e$; $z \leq 18.5R_e$; $\sqrt{x^2+y^2} \leq 11.6R_e$; $\sqrt{x^2+z^2} \leq 19.8R_e$; $\sqrt{z^2+y^2} \leq 20.0R_e$. These constraints were chosen by rounding the maxima and minima of the spatial parameters to the nearest tenth. As a result, we were left with a trimmed dataset, reduced from 1,172,923 to 462,615 data points. Though compact, this dataset is centered on the space region significant for SMILE, promising heightened model accuracy. It's noteworthy that negative z-values weren't excluded, considering the SMILE mission's highly inclined, elliptical orbit, which dips to -3.5 R_e. Omitting these would disregard vital data, especially since the Cluster's trajectory spent a notable amount of time in the southern hemisphere.

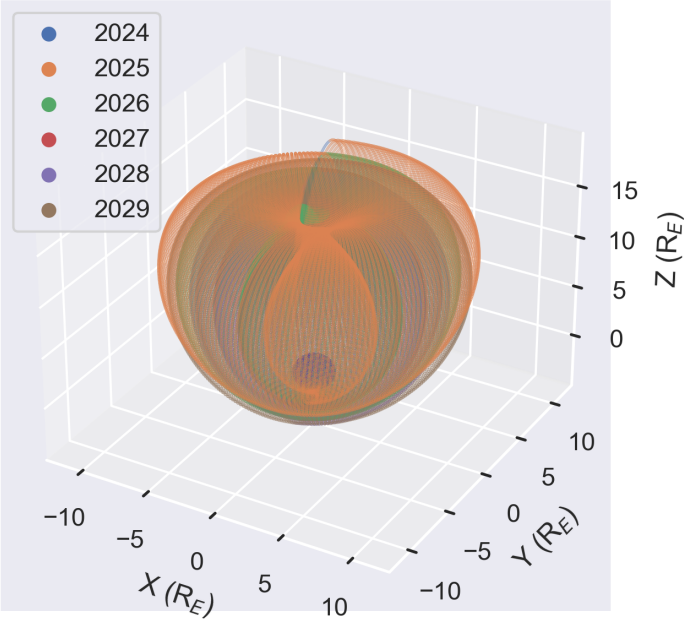


Figure 1. SMILE mission's trajectory. Distinct colors represent individual years with Earth, having a radius of 6,731 km, at the center.

2.2 Predictor Introduction

The spacecraft's location in the Geocentric Solar Ecliptic (GSE) coordinate system is defined by `x`, `y`, and `z` in Earth radii (R_e). The variable `rdist` denotes the satellite's radial distance from Earth. The magnetic field line type, termed as "Foot Type", indicates the connectivity of the IMF field lines to Earth, calculated using the Tsyganenko (1995) model. The initial definition stated by Kronberg et al. (2020) is as follows: the interplanetary magnetic field lines (IMF) with no connection to Earth have Foot Type 0, open magnetic field lines with one connection to Earth have Foot Type 1, and closed field lines with both ends connected to Earth have Foot Type 2. It was, however, decided to redefine the IMF to 1 and open field lines to 0, to achieve a stronger linear relationship between Foot Type and the target variable, as discussed in chapter 2.3.2.

The Disturbance storm time index (`Dst_index`) characterizes geomagnetic storms in the unit nT (Banerjee et al., 2012). The Auroral Electrojet (`AE_index`) quantifies magnetic activity in the auroral zone, also denoted in nT. The 10.7 cm solar radio flux (F10.7) with unit sfu serves as a solar activity level indicator and a proxy for solar emissions (Tapping, 2013).

The IMF direction is described by its components `BimfxGSE`, `BimfyGSE`, and `BimfzGSE` in the GSE system in nT. The IMF direction at the magnetopause determines if reconnection happens on the dayside (Crooker et al., 1979; Luhmann et al., 1984; Koga et al., 2019). Plasma properties of the solar wind are described by Solar wind speed (`VSW`) in km/s, proton density (`NpSW`) in cm^{-3}, and temperature (`Temp`) in K. The direction of the solar wind velocity is described by `VxSW_GSE`, `VySW_GSE`, and `VzSW_GSE`. The solar wind dynamic pressure, `Pdyn` (nPa), can be represented as:

$$\texttt{Pdyn} = \texttt{NpSW} * \texttt{VSW}^2 * 1.67 * 10^{-6} \quad (1)$$

2.3 Exploratory Data Analysis

2.3.1 Spatial Proton Intensity Distribution

To analyze the proton intensity distribution in relation to the spacecraft's trajectory, we combined the y and z coordinates to produce a radial distance variable, termed `yz_axis`. We introduced this variable because Cluster's trajectory is predominantly in the southern hemisphere, contrasted with SMILE's expected northern trajectory. The `yz_axis` is computed by $\sqrt{y^2 + z^2}$, offering a simplified yet informative perspective on proton intensity's spatial distribution. Figure 2 depicts the spatial distribution of proton intensities in the x, $\sqrt{y^2 + z^2}$ coordinate system. The color gradient represents the percentage of measurements that exceed 2, the mean value of $\log_{10}$(proton intensities) rounded to one significant digit, highlighting regions with prolonged high proton intensities. The central black void indicates missing measurements. This gap arises from our deliberate exclusion of data points with radial distances (`rdist`) below 6 Earth radii (R_e) in order to emphasize regions beyond the radiation belts. Historically, proton intensities surge in zones under 6 R_e, which encompass the ring current and radiation belt regions. Our focus shifts to lesser-studied areas, with their generally lower intensities outside of the radiation belts, mainly because the SXI telescope on the SMILE mission is equipped with a shutter mechanism, protecting its Charge Coupled Devices (CCD) from intense radiation within the radiation belt. The decline in proton intensities with increased distance from Earth is observed, irrespective of whether it's along the `x` axis or the y-z plane. This observation aligns with subsequent feature plots and correlation matrix analyses. Moreover, this analysis reveals areas along closed magnetic field lines with heightened proton intensities, as well as sparser regions corresponding to open magnetic field lines over the polar cap, demonstrating a clear spatial correlation between magnetic field line configuration and proton intensity distribution.

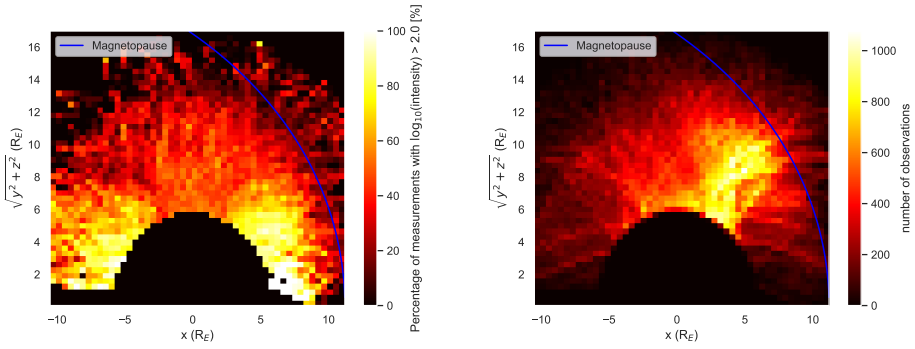


Figure 2. (Left) Heatmap of proton intensities against the x-coordinate and $\sqrt{y^2 + z^2}$. The color gradient represents the percentage of measurements where $\log_{10}$(proton intensity) > 2, which is the mean value rounded to one significant digit. (Right) Data point density for each bin, with the blue line representing the magnetopause, derived using Shue et al. (1997). Both plots incorporate the 462,615 data points post-reshaping for the SMILE mission.

2.3.2 Cross-Correlation and Feature Plot Analysis

Cross-correlation matrices, employing the Pearson coefficient, are used in feature selection for linear regression models. The coefficient quantifies the linear relationship

strength and direction between two variables, spanning from -1 (perfect negative relationship) to 1 (perfect positive relationship), with 0 indicating no linear correlation. Such analyses illuminate potential multicollinearity issues in datasets, which can adversely affect regression coefficient stability and model interpretability (Raschka et al., 2022; James et al., 2013).

Analyzing the cross-correlation matrix (Figure 3), we observed:

- `FootType`: The feature plot in Figure 4 highlighted a clear potential for refining the correlation between `FootType` and `p3`. The initial positive correlation of 0.24 with `p3` was improved to 0.41 upon redefining the foot type as mentioned in section 2.2.
- `AE_index`: While the correlation was weak (0.07), the feature plot identified AE values surpassing 2600 nT as possible outliers.
- F10.7 solar radio flux index: Given its correlation coefficient of -0.20, the feature plot shows a predominantly monotonically decreasing relationship between F10.7 and the target variable.
- `VxSW_GSE`: The feature plot demonstrated that proton intensity increases with higher absolute wind speeds up to 950 km/s. Values exceeding this were considered as potential outliers.
- Distance variables: While `z` showed a positive correlation of 0.21, its relationship with proton intensities displayed a clear maximum around 0 on the feature plot. This insight led to the introduction of `|z|` as an improved predictor.
- Other Variables: The strong negative correlations of `rdist` and `yz_axis` with `p3` were supported by the feature plot's linear regression lines, emphasizing their importance as predictors.

Analysis of the proton intensity histogram identified two extreme outliers exceeding 100,000 $1/cm^2/s/sr/keV$ (see Figure A1). Removing these and other above-identified outliers from different predictors did not improve model performance, justifying their retention. Further analysis revealed 606 F10.7 measurements above 900, deemed unrealistic and consequently removed.

2.4 Data Split and Data Scaling

This section outlines the additional processing steps applied to the data set, already reshaped for the SMILE mission as detailed in section 2.1. These steps include splitting the data into training and testing sets, transforming the target variable, and scaling features.

Records before December 31, 2012, were allocated to the training set, while records from January 1, 2013, onwards formed the test set. This temporal division results in an approximate 75% to 25% split between the training and test datasets. The training set, was then later further divided into training and validation sets by the use of five-fold cross-validation, where the dataset is divided into five parts, with each part being used as a validation set while the remaining four parts are used as training data

The proton intensities recorded in channel 3 (`p3`), our target values, display a wide spectrum. We therefore transformed these values using a base 10 logarithmic function. Addressing the challenge of logging zero values, all such occurrences in `p3` were substituted with 0.5. However, this introduces potential pitfalls as we expect an artificial population with the same values, a concern later revisited during model evaluation (Bellégo et al., 2021).

Optimizing gradient descent requires careful attention to feature scaling (Raschka et al., 2022). In our polynomial regression model, we employed a double-scaling tech-

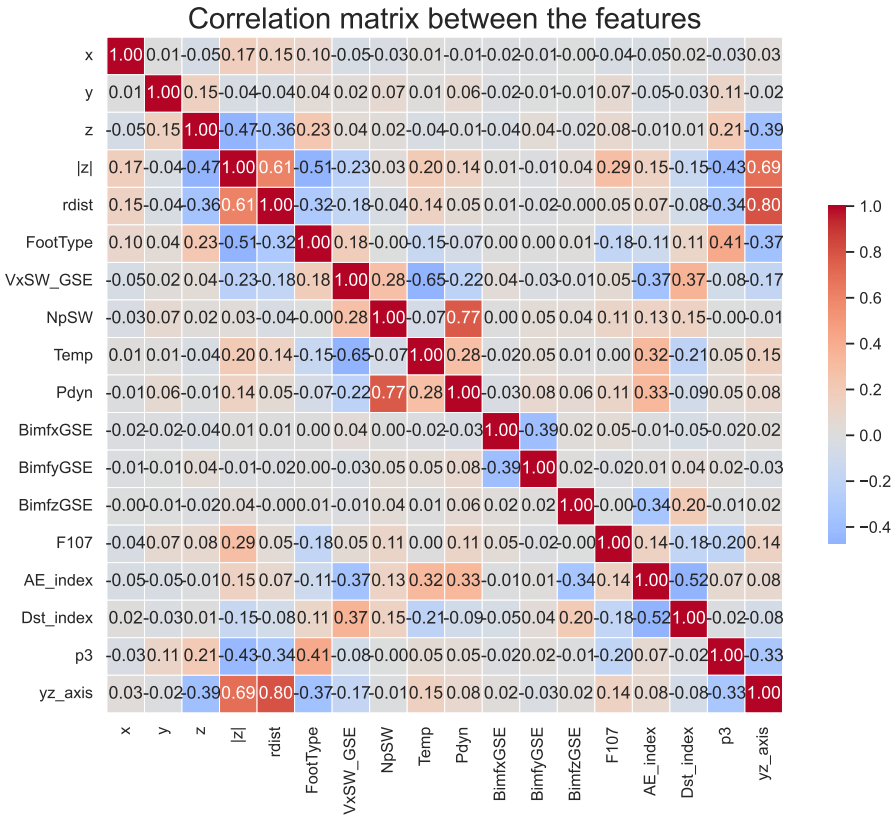


Figure 3. Pearson coefficient-based correlation matrix for the predictors and the proton intensity post-data reshaping for the SMILE mission.

nique to ensure numerical stability and facilitate the optimization process. Initially, the original features were scaled to zero mean and unit variance using the `StandardScaler()` method, aligning with the desired outcomes (Pedregosa et al., 2011). Subsequently, polynomial features were generated from these scaled features. To further enhance the model's robustness, these polynomial features were subjected to a second round of scaling using the same `StandardScaler()` method.

By scaling both the original and polynomial features, we ensure that the coefficients are directly comparable in terms of their contribution to the model and that all features display a mean and unit variance of zero.

3 Methodology

3.1 Linear Regression Model

The choice of employing linear regression models in this study is underpinned by several reasons. First and foremost, linear regression models offer a simple and interpretable framework for understanding how input variables affect the output. Furthermore, the methodology allows for the transformation of input variables to enhance their predictive capabilities, such as the introduction of polynomial terms and interaction effects (Hastie et al., 2001).

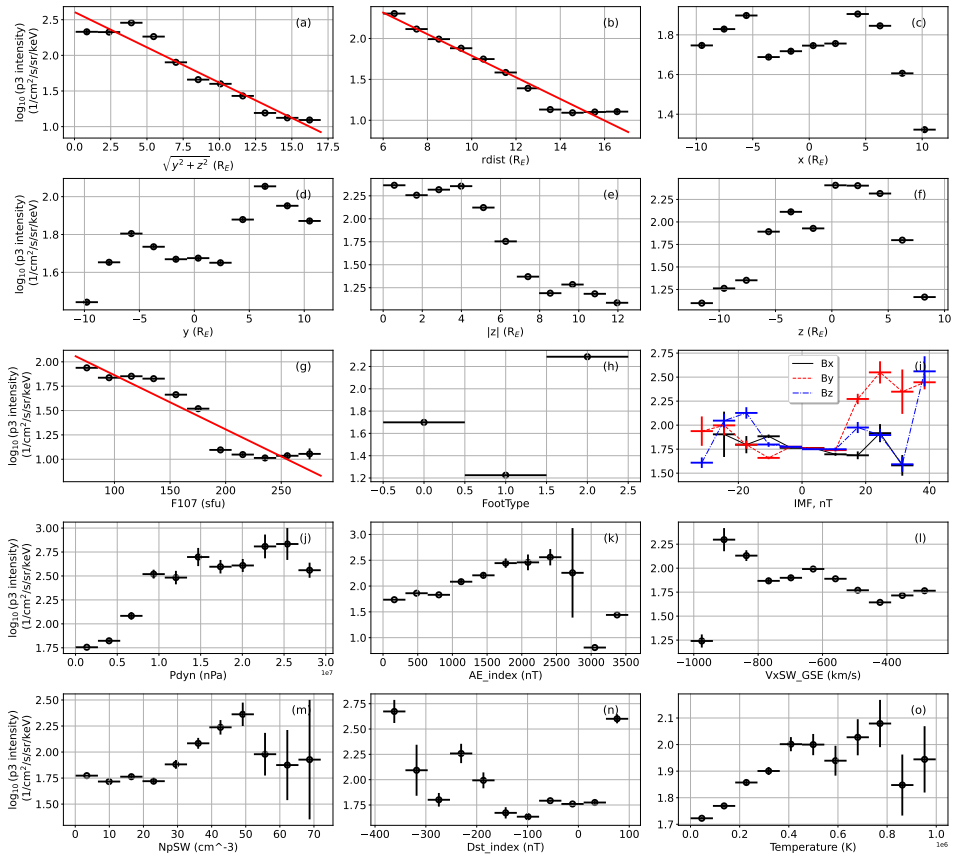


Figure 4. Mean of the logarithmically scaled proton intensities from the `p3` channel against potential predictors. Vertical lines depict the standard 95% confidence level, while horizontal lines indicate bin half-widths. Linear regression lines in red are shown for `rdist`, `yz_axis`, and F10.7.

The Ordinary Least Squares (OLS) model serves as the foundational approach, focusing on minimizing the sum of squared differences between observed and predicted values (James et al., 2013; Galton, 1886). To tackle the possible issue of multicollinearity, Ridge Regression can be utilized, which incorporates an L2 penalty term into the loss function (Kutner et al., 2005; Hoerl & Kennard, 1970). Lasso Regression is employed when feature selection is essential, as it uses an L1 penalty to drive certain coefficients to zero, effectively eliminating them from the model (Santosa & Symes, 1986). Lastly, Elastic Net Regression can be used to combine the strengths of both L1 and L2 penalties, providing a balanced approach that can handle both multicollinearity and feature selection (Pedregosa et al., 2011). Multiple models were trained using the scikit-learn library in Python (Pedregosa et al., 2011).

3.2 Model Selection and Optimization

For model evaluation, we utilized a set of metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2 (coefficient of determination), Pearson correlation, and Spearman correlation. Model selection was primarily guided by the performance of R^2 and Spearman correlation on the validation set. To ensure a robust and generalizable evaluation, five-fold cross-validation with the help of the `KFold` function from `sklearn.model_selection` was applied to the training set. Given the time-series nature of our dataset, the `shuffle` parameter within the cross-validation procedure was intentionally set to `false`. Subsequently, the evaluation metrics were computed as the average values derived from the five cross-validation folds, thereby offering a more reliable measure of the model's true performance.

3.2.1 Simple OLS, Lasso, Ridge and Elastic Net

Following the initial selection of linear regression models, two distinct approaches were undertaken to optimize model performance. The first approach involved the application of various linear regression techniques, including OLS, Lasso, Ridge, and Elastic Net. This approach, however, did not yield satisfactory results. The maximum R^2 value on the validation set was only 0.02, and the highest Spearman correlation coefficient was 0.43.

3.2.2 Introduction of Polynomial Terms

To improve upon this, the second approach incorporated polynomial terms into a standard Lasso model from `sklearn.linear`. The model was optimized for the regularization parameter α using five-fold cross-validation. The optimal α was determined using `LassoCV` with a maximum iteration of 10,000 and a tolerance of 1×10^{-5}. This approach significantly improved the model performance, achieving an R^2 value of 0.22 and a Spearman correlation coefficient of 0.51 on the validation set.

3.2.3 Heuristic-based Feature Selection Technique

However, this model included 52 predictors, making it complex and potentially prone to overfitting. Further work was needed to develop a more parsimonious model with a maximum of 25 predictors while maintaining acceptable performance. To reduce the number of predictors while maintaining model performance, we adopted a heuristic-based feature selection strategy. For this strategy the Stochastic Gradient Descent (SGD) framework was employed, with the algorithm configured as follows: the regularization term (α) was set to the optimal value identified through cross-validation. The learning rate was set to a constant value, initialized at $\eta_0 = 1 \times 10^{-5}$. The hyperparameter defining the loss function was set to the squared error loss. An L1 penalty term was incor-

porated for feature selection. The algorithm was set to terminate when the tolerance reached 1×10^{-5}, with a maximum of 100 iterations for convergence.

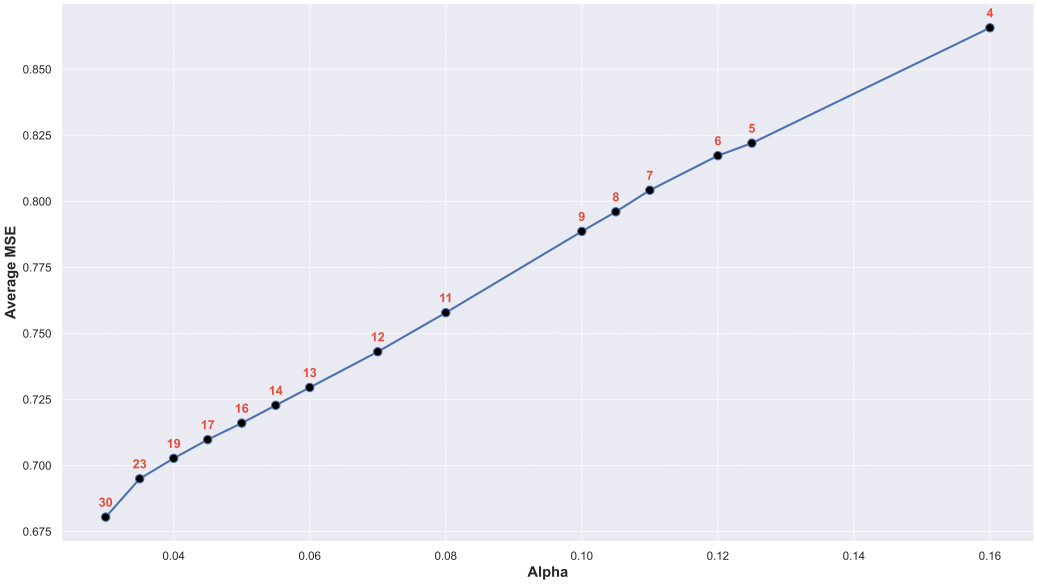


Figure 5. Plot of Average Mean Squared Error (MSE) against the regularization parameter α. The curve exhibits an "elbow" point at 13 predictors, indicating a minimal but acceptable loss in model performance. A noticeable increase in MSE is observed when the number of predictors is reduced from 13 to 12, suggesting that all 13 predictors left, display significant importance for the model. This "elbow" point serves as the basis for selecting an optimal α value and, consequently, the number of predictors for the final model.

Unlike earlier approaches that solely aimed to minimize the Mean Squared Error (MSE), this method also considers the number of predictors in the final model. We tested a range of regularization parameters (α) and sought to identify a "knee" or "elbow" in the plot of MSE versus α. This point represents a compromise between model performance and complexity.

To enhance the robustness of the feature selection process, we employed K-Fold cross-validation with the shuffle parameter set to `True`. This approach allows for a more representative sampling of the training data across each fold. Specifically, we aimed to identify the most stable set of predictors corresponding to the "elbow" point for the regularization parameter α. By enabling shuffling during cross-validation, we increase the likelihood that the predictor set extracted from one of the folds offers a more comprehensive representation of the entire training dataset. The α range chosen was from 0.03 to 0.17, which covered all models with the amount of non-zero predictors ranging from 28 to 4.

Upon employing this approach, we identified a subset of 13 predictors by analyzing the MSE vs α plot in Figure 5. Importantly, we operate under the assumption that all predictors remaining after the feature selection process are relevant to the outcome. Therefore, penalizing these predictors, as Lasso does, could introduce an unwanted bias into the model. Given this consideration, an OLS model was chosen for the final training rather than a Lasso regression model.

Table 1. Average performance metrics for different models resulting from five-fold cross-validation and the number of data points N_train used for training.

Model	MSE	MAE	R^2	Pearson	Spearman	Predictors	N_train
Basic_OLS	0.92	0.76	0.02	0.43	0.43	9	353,660
Poly_Lasso	0.73	0.68	0.22	0.51	0.51	52	353,660
Heuristic_Poly_OLS	0.78	0.71	0.17	0.47	0.47	13	353,660
Split_Poly_Part1	0.58	0.60	0.24	0.53	0.53	5	60,961
Split_Poly_Part2	0.82	0.73	0.09	0.50	0.50	19	145,952
Split_Poly_Part3	0.79	0.72	0.19	0.46	0.46	15	70,137
Split_Poly_Part4	0.65	0.64	0.29	0.55	0.55	6	76,610

Although the resulting model exhibits lower performance on the validation set, as evidenced by Table 1, it better aligns with the study's objectives of interpretability and usability compared to the Lasso model with 52 predictors. This heuristic-based feature selection strategy aligns well with the principle of Occam's razor, suggesting that simpler models are preferable when performance is comparable. Therefore, this approach effectively strikes a balance between the number of predictors and model performance, thereby enhancing the model's interpretability and practical utility.

3.2.4 Data Split

An in-depth analysis of the relationship between the `y` and `p3` variables revealed that the data could be divided into four distinct parts, each characterized by an increasing or decreasing slope, see Figure 4 (d). This lead to the decision to split the dataset into four separate parts based on specific conditions, as described below:

- Part 1: $\mathtt{y} \leq$ -6.6 R_e
- Part 2: -6.6 $R_e \leq \mathtt{y} \leq$ 2.3 R_e
- Part 3: 2.3 $R_e \leq \mathtt{y} \leq$ 6 R_e
- Part 4: $\mathtt{y} \geq$ 6 R_e

Upon splitting the data, separate models were built for each part, using the same heuristic-based predictor selection technique previously described. These outperformed the non-split OLS model in the Spearman correlation coefficient and the R^2 metric for three out of the four subsets (see Table 1), all while maintaining low model complexity.

4 Results

This chapter presents the empirical results obtained from the evaluation of various OLS models on the unseen test set. The models are compared based on a set of evaluation metrics and feature importances.

4.1 Presentation of Final Models

In this section, we present the final forms of our linear regression models developed for predicting proton intensities. Each model is displayed with its coefficients in basic, unscaled units, offering a clear view of the relative impact of each predictor variable. These models encapsulate our findings and are ready for practical application.

Heuristic_Poly_OLS:

$$\begin{aligned}\log_{10}(p3) = {} & 2.1 \times 10^{-2} \cdot y - 3.0 \times 10^{-1} \cdot |z| - 1.7 \times 10^{-1} \cdot \text{rdist} + 1.1 \cdot \text{FootType} \\ & - 2.0 \times 10^{-3} \cdot \text{VxSW\_GSE} + 6.2 \times 10^{-8} \cdot \text{Pdyn} - 1.1 \times 10^{-3} \cdot \text{F107} \\ & + 3.7 \times 10^{-4} \cdot \text{AE\_index} + 8.3 \times 10^{-4} \cdot x \cdot |z| + 1.8 \times 10^{-2} \cdot |z| \cdot \text{rdist} \\ & - 7.4 \times 10^{-2} \cdot \text{rdist} \cdot \text{FootType} + 7.7 \times 10^{-4} \cdot \text{rdist} \cdot \text{F107} \\ & - 3.1 \times 10^{-5} \cdot \text{F107}^2 + 2.3 \end{aligned} \tag{2}$$

Split_Poly_Part1:

$$\begin{aligned}\log_{10}(p3) = {} & -5.9 \times 10^{-1} \cdot \text{rdist} + 8.4 \times 10^{-1} \cdot \text{FootType} - 2.5 \times 10^{-3} \cdot \text{VxSW\_GSE} \\ & + 2.0 \times 10^{-2} \cdot \text{rdist}^2 - 5.0 \times 10^{-2} \cdot \text{rdist} \cdot \text{FootType} + 4.2 \end{aligned} \tag{3}$$

Split_Poly_Part2:

$$\begin{aligned}\log_{10}(p3) = {} & -6.2 \times 10^{-1} \cdot |z| + 2.3 \times 10^{-1} \cdot \text{rdist} + 7.5 \times 10^{-1} \cdot \text{FootType} \\ & + 7.8 \times 10^{-4} \cdot \text{VxSW\_GSE} + 1.3 \times 10^{-7} \cdot \text{Pdyn} + 3.0 \times 10^{-4} \cdot \text{AE\_index} \\ & + 3.4 \times 10^{-3} \cdot x \cdot |z| + 3.4 \times 10^{-5} \cdot x \cdot \text{VxSW\_GSE} - 5.3 \times 10^{-4} \cdot y \cdot \text{rdist} \\ & + 5.1 \times 10^{-2} \cdot |z| \cdot \text{rdist} - 3.7 \times 10^{-2} \cdot \text{rdist}^2 - 7.2 \times 10^{-2} \cdot \text{rdist} \cdot \text{FootType} \\ & - 2.5 \times 10^{-4} \cdot \text{rdist} \cdot \text{VxSW\_GSE} + 6.0 \times 10^{-4} \cdot \text{rdist} \cdot \text{F107} \\ & + 1.8 \times 10^{-1} \cdot \text{FootType}^2 - 5.7 \times 10^{-8} \cdot \text{FootType} \cdot \text{Pdyn} \\ & + 9.2 \times 10^{-14} \cdot \text{Temp} \cdot \text{Pdyn} \quad - 1.9 \times 10^{-16} \cdot \text{Pdyn}^2 - 2.3 \times 10^{-5} \cdot \text{F107}^2 + 1.9 \end{aligned} \tag{4}$$

Split_Poly_Part3:

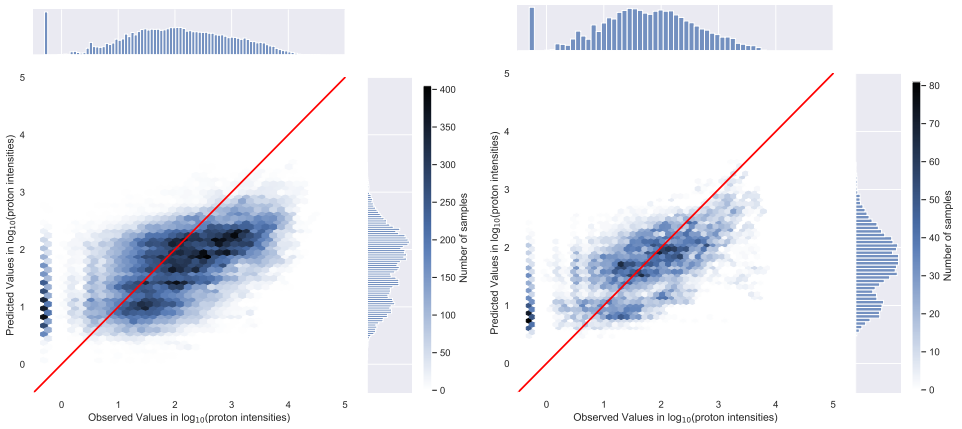
$$\begin{aligned}\log_{10}(p3) = {} & 3.6 \times 10^{-1} \cdot y - 2.8 \times 10^{-1} \cdot |z| - 5.7 \times 10^{-2} \cdot \text{rdist} + 1.1 \cdot \text{FootType} \\ & - 1.7 \times 10^{-3} \cdot \text{VxSW\_GSE} + 2.5 \times 10^{-7} \cdot \text{Temp} + 5.2 \times 10^{-8} \cdot \text{Pdyn} \\ & - 1.3 \times 10^{-3} \cdot \text{F107} + 4.6 \times 10^{-4} \cdot \text{AE\_index} + 2.5 \times 10^{-3} \cdot x \cdot |z| \\ & - 2.8 \times 10^{-2} \cdot y \cdot \text{rdist} + 1.4 \times 10^{-2} \cdot |z| \cdot \text{rdist} - 8.2 \times 10^{-2} \cdot \text{rdist} \cdot \text{FootType} \\ & + 8.2 \times 10^{-4} \cdot \text{rdist} \cdot \text{F107} - 3.1 \times 10^{-5} \cdot \text{F107}^2 + 1.2 \end{aligned} \tag{5}$$

Split_Poly_Part4:

$$\begin{aligned}\log_{10}(p3) = {} & -5.1 \times 10^{-2} \cdot |z| - 1.6 \times 10^{-1} \cdot \text{rdist} + 2.2 \times 10^{-1} \cdot \text{FootType} \\ & - 2.6 \times 10^{-3} \cdot \text{VxSW\_GSE} - 1.9 \times 10^{-3} \cdot \text{F107} + 7.4 \times 10^{-4} \cdot \text{AE\_index} + 2.7 \end{aligned} \tag{6}$$

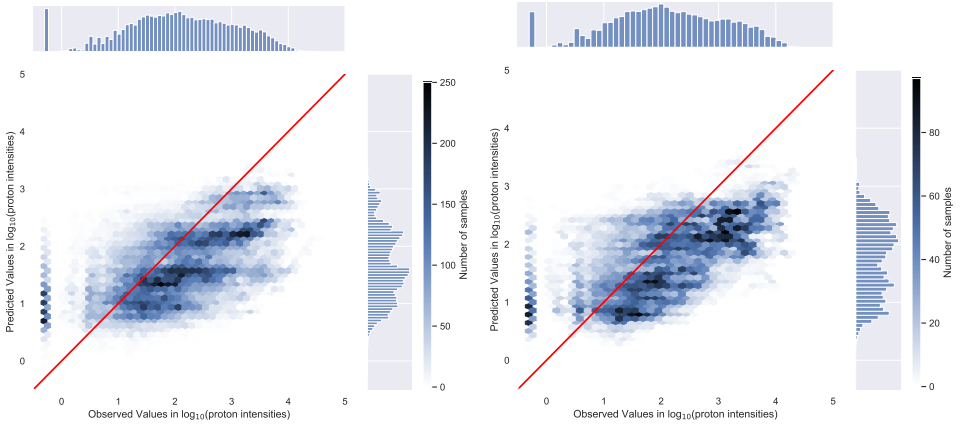
4.2 Performance on the Test Set

In the Heuristic_Poly_OLS model, the hexagonal bins largely align with the ideal fit line (see Figure 6), which is indicative of good predictive performance. However, this model exhibits a tendency to underestimate observed values, notably at higher proton intensities. A significant peak at log(0.5) in the histogram of observed values is associated with an overestimation in the Heuristic_Poly_OLS model's predictions. This peak stems from the substitution of zero values in the target variable $p3$ before applying the logarithmic transformation. This overestimation at log(0.5) potentially skews the model's learning process, causing it to adjust its predictions downward to minimize the overall loss. While this adjustment mitigates the error for overestimated values, it concurrently introduces a bias leading to the underestimation of other observed values. This behav-



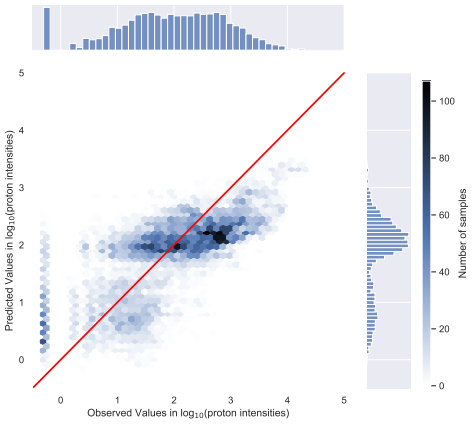
(a) Heuristic_Poly_OLS model

(b) Split_Poly_Part1



(c) Split_Poly_Part2

(d) Split_Poly_Part3



(e) Split_Poly_Part4

Figure 6. Jointplots comparing observed and predicted values of proton intensities, of the test set for the different OLS models. The red lines represent ideal fits where observed values equal predicted values. Color bars indicate the number of samples in each hexagonal bin. Histograms at the top and right margins show the distributions of observed and predicted values for each model.

Table 2. Performance metrics of the final models on test data.

Model	MSE	MAE	R^2	Pearson	Spearman	Predictors
Heuristic_Poly_OLS	0.74	0.71	0.22	0.56	0.57	13
Split_Poly_Part1	0.46	0.54	0.38	0.62	0.61	5
Split_Poly_Part2	0.83	0.74	0.11	0.56	0.57	19
Split_Poly_Part3	0.77	0.73	0.24	0.61	0.62	15
Split_Poly_Part4	0.47	0.56	0.50	0.72	0.72	6

ior is consistently observable across all models, particularly those focusing on specific regions. As an alternative to zero substitution, we also explored the removal of these zero values. While this approach enhanced performance on the training set, it consistently led to diminished performance on the validation set. Consequently, despite its limitations, the zero substitution technique was retained to ensure better generalization to unseen data.

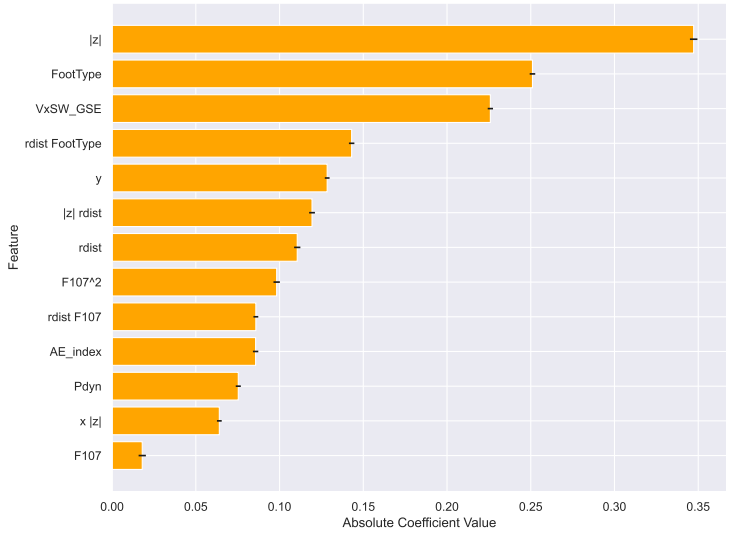
Turning our attention to Split_Poly_Part1 and Split_Poly_Part4, these models exhibit the most well-centered distribution around the ideal fit line in their respective heatmaps. This observation is consistent with their performance metrics as recorded in Table 2, showcasing R^2 values of 0.38 and 0.50 and Spearman coefficients of 0.61 and 0.72 for the test set. Notably, these models also maintain low complexity, employing only 5 and 6 predictors, respectively.

Conversely, Split_Poly_Part2, with its high complexity due to having 19 predictors, exhibits subpar performance despite an acceptable Spearman coefficient. Significantly, with an R^2 value of only 0.11, this model is the sole split variant that exhibits notably inferior performance compared to the unsplit Heuristic_Poly_OLS model in the test set. The more inhomogeneous distribution of hexagonal bins in its heatmap is indicative of this weaker performance. On the other hand, Split_Poly_Part3 shows a modest improvement over the unsplit model. This is evident not only in the performance metrics but also in a more concentrated distribution in its heatmap, compared to Split_Poly_Part2.

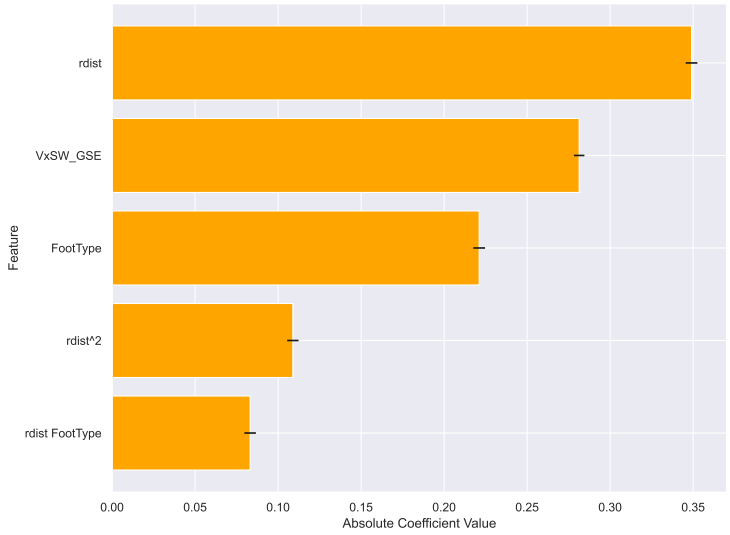
4.3 Feature Importance

In order to derive feature importance in a linear regression model, one can examine the coefficients of the model. The magnitude of the coefficients indicates the relative importance of the corresponding feature in predicting the target variable. A larger absolute value of a coefficient suggests a stronger influence of the associated feature on the outcome. The features are scaled appropriately by scaling the features once before the creation of the polynomials and once afterward. Scaling ensures that all features are on a comparable scale, which prevents features with larger values from dominating those with smaller values in the model. The feature importance for each model was plotted in figure 7.

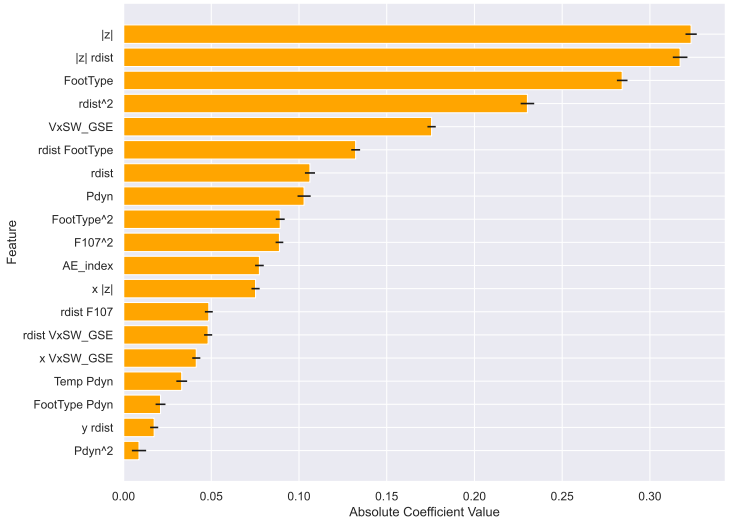
The variations in feature importance across the different models offer insights into the underlying mechanisms affecting proton intensities in various regions. For the Heuristic_Poly_OLS model and models corresponding to the inner regions (Split_Poly_-Part2 and Split_Poly_Part3), the absolute value of z (`|z|`) emerges as the most significant predictor, next to `Foottype` and `VxSW_GSE`. The Split_Poly_-Part2 model additionally identifies the polynomial terms `|z| rdist` and $\texttt{rdist}^2$ as significant features. The similarity between the models for the inner part and the model trained on the full data is most likely partially influenced by the fact that the inner regions contain 61% of the total data points. The models tailored to the outer regions (Split_Poly_Part1 and Split_Poly_Part4)



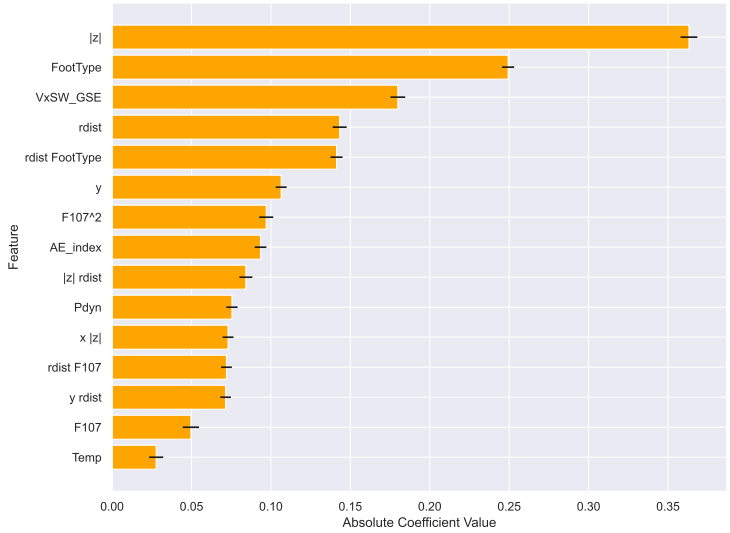
(a) Heuristic_Poly_OLS model



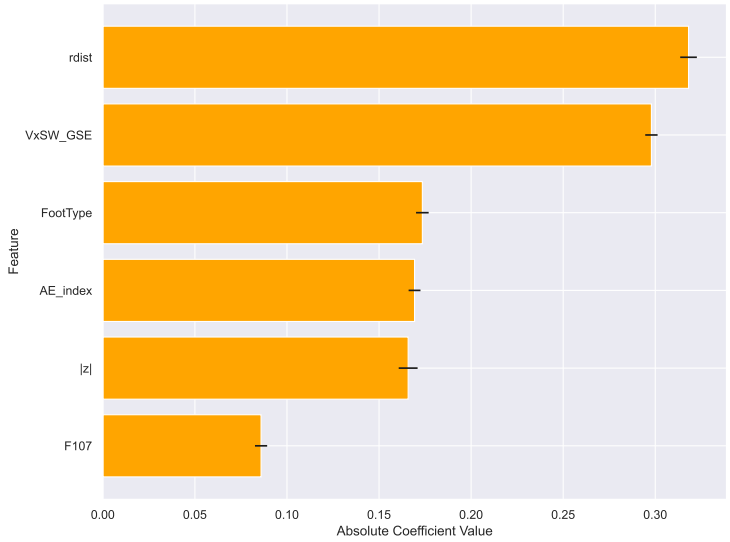
(b) Split_Poly_Part1



(c) Split_Poly_Part2



(d) Split_Poly_Part3



(e) Split_Poly_Part4

Figure 7. Feature importance plots for five different OLS models: Each plot presents the absolute values of the model coefficients, serving as indicators of feature importance. Accompanying error bars represent the standard errors, providing a measure of the coefficient's reliability. The plots collectively offer insights into the relative significance of each predictor across different models.

prioritize `rdist`, `VxSW_GSE`, and `Foottype` as their top predictors, in that order. A notable distinction from the inner region models is the elevated significance of `VxSW_GSE`.

5 Discussion

The most critical predictor for our Heuristic_Poly_OLS_Model, which utilizes the full dataset, is the absolute value of z, denoted as $|z|$. The model reveals a negative correlation between $|z|$ and the proton intensities, indicating that as $|z|$ increases, the proton intensity declines. This trend can be primarily attributed to the circulation of protons in Earth's magnetic field. Most ions are concentrated at the equatorial plane during their drift trajectories on the closed magnetic field lines. At higher latitudes where open magnetic field lines dominate, the proton intensities are expected to drop with $|z|$ distance. Consequently, the proton intensities reduce with an increase in $|z|$.

The predictor `FootType` categorizes magnetic field line types and ranks as the second most influential factor. Closed field lines, known for the highest proton intensities, trap charged particle populations. The importance of this parameter aligns with the studies by Walsh et al. (2014) and Kronberg et al. (2020). In contrast, open field line regions typically correlate with lower particle energies outside the soft proton (SP) range, resulting in weaker count rates as detailed in (Kronberg et al., 2020). IMF regions show slightly higher count rates since particles can experience acceleration in the bow shock region, especially quasi-parallel bow shock configurations (normal to the shock is parallel to the IMF direction) (Blandford & Ostriker, 1978; Kronberg et al., 2009; Sundberg et al., 2016).

The high importance of solar wind speed in the X-direction is consistent with the analysis of the feature plot in figure 4. Kronberg, Hannan, et al. (2021) also found that `VxSW_GSE` displays "the most substantial linear dependence of the proton intensities among the OMNI parameters." The solar wind speed, directly correlated to its electric field as $E = V_x \times B_z$, is crucial for magnetospheric dynamics as it determines the rate of magnetic reconnection on Earth's dayside (Dorelli, 2019), and consequently magnetic reconnection at the night side. A surge in solar wind speed correlates with an increased rate of magnetic reconnection. Additionally, magnetic reconnection events, which can accelerate charged particles, also impact soft proton intensities significantly, as noted by (Read & Ponman, 2003). Research by Gonzalez et al. (1994), Milan et al. (2012)), and Wang et al. (2014) further elucidates this concept, indicating that a variety of solar wind-magnetosphere energy transfer models are dependent on the velocity of the solar wind.

6 Conclusion

In this study, we developed five user-friendly linear regression models to predict proton intensities in the energy range of 92.2 keV to 159.7 keV with a Spearmen correlation ranging from 0.57 to 0.72 on the test data. Utilizing data from the Cluster's RAPID experiment, supplemented with solar, solar wind, and geomagnetic data from the OMNI database, the study focused on aligning the models with the anticipated spatial area covered by the upcoming SMILE-mission.

Segmenting the data into four distinct regions based on the y coordinate with thresholds -6.6 R_E, 2.3 R_E and 6 R_E, resulted in enhanced model performance for three of the four segments, surpassing the main model's performance. The primary predictors in these outer regions were identified as radial distance and the radial solar wind speed. Conversely, the inner region models and the comprehensive main model demonstrated a significant dependence on the absolute value of z and the type of magnetic field lines.

The redefinition of the `FootType` variable and the incorporation of the absolute value of z as key model features significantly improved the model compared to previous relevant studies. This study suggests that the development of more accurate predictive mod-

els for space weather phenomena may not solely rely on novel algorithms, but also on crafting tailored models, each addressing distinct regions with their specific characteristics.

Appendix A : Histogram of Proton Intensities

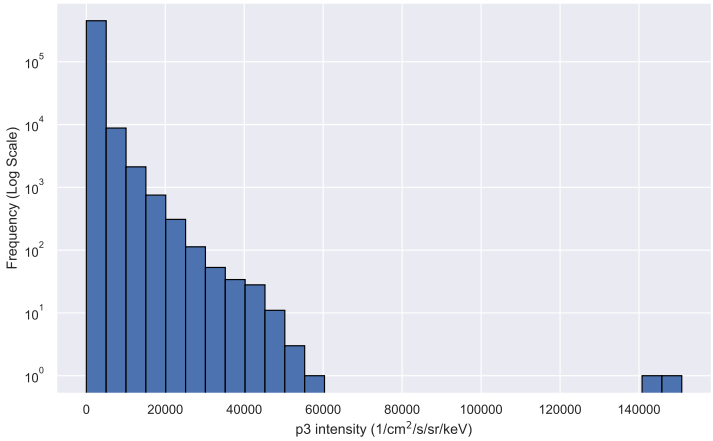


Figure A1. Histogram of the proton intensities measured by channel 3.

Appendix B Open Research

The authors express their gratitude to the team at the Cluster Science Archive (`https://csa.esac.esa.int`) for supplying the data. Additionally, we recognize the utilization of the OMNIWeb service and OMNI data from NASA/GSFC's Space Physics Data Facility (King & Papitashvili, 2005). The code and dataset used to derive the linear regression model can be found via the following link: `https://zenodo.org/records/10964236?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjkxMGQzY2EzLTMxNDAtNDZmNi05MjE5LWUzZmM1Y2I2OWM5MSIsImRhdGEiOnt9LCJyYW5kb20iOiIyMjQ1NjQzYTdlZjk3YzNjODA1MzdlZGJhMmQyMzg2MyJ9.oKgcCJTFE6KbvqqjXNh3wzfneL3XeY6meWb-XhbcKumOx0ztugSGFtvCaLblb3WAWOE5ccrkVEWZDnZ9vEs6EQ`.

Acknowledgments

The database used in this study was generated within the team led by Fabio Gastaldello on "Soft Protons in the Magnetosphere focused by X-ray Telescopes" at the International Space Science Institute in Bern, Switzerland. We are particularly grateful to Dr. Gastaldello for his invaluable feedback on our work, which greatly enhanced this research. EK and SM are supported by the German Research Foundation (DFG) under number KR 4375/2-1 within SPP "Dynamic Earth". EK is also supported by the DFG under number KR 4375/4-1. We are grateful to Dr. Andrew Read and Dr. Steven Sembay for their insightful suggestions which significantly enhanced the representation of spatial proton intensity distribution in our plots.

References

Banerjee, A., Bej, A., & Chatterjee, T. N. (2012). On the existence of a long range correlation in the geomagnetic disturbance storm time (dst) index. *Astrophysics and Space Science*, *337*, 23–32. doi: 10.1007/s10509-011-0836-1

Bellégo, C., Benatia, D., & Pape, L. (2021). Dealing with logs and zeros in regression models. *CREST - Serie des Documents de Travail*. doi: 10 .2139 / ssrn .3444996

Blandford, R. D., & Ostriker, J. P. (1978, April). Particle acceleration by astrophysical shocks. *Astrophys. J.*, *221*, L29-L32. doi: 10.1086/182658

Branduardi-Raymont, G., & Wang, C. (2022). The smile mission. In C. Bambi & A. Santangelo (Eds.), *Handbook of x-ray and gamma-ray astrophysics* (pp. 1–22). Singapore: Springer Nature Singapore. doi: 10.1007/978-981-16-4544-0_39-1

Branduardi-Raymont, G., Wang, C., Escoubet, C. P., Adamovic, M., Agnolon, D., Berthomier, M., ... Zhu, Z. (2018). *Smile definition study report* (ESA/SCI No. 1). European Space Agency. doi: 10.5270/esa.smile.definition_study_report-2018-12

Crooker, N. U., Eastman, T. E., & Stiles, G. S. (1979). Observations of plasma depletion in the magnetosheath at the dayside magnetopause. *Journal of Geophysical Research: Space Physics*, *84*(A3), 869-874. doi: 10.1029/JA084iA03p00869

Daly, P. W., & Kronberg, E. A. (2023). *User guide to the rapid measurements in the cluster science archive (csa)* (User Guide No. CAA-EST-UG-RAP 6.1). Max Planck Institute for Solar System Research. Retrieved 2023-11-05, from `https://www2.mps.mpg.de/dokumente/projekte/cluster/rapid/Rapid_Userguide.pdf`

Dorelli, J. C. (2019). Does the solar wind electric field control the reconnection rate at earth's subsolar magnetopause? *Journal of Geophysical Research: Space Physics*, *124*(4), 2668-2681. doi: 10.1029/2018JA025868

Fioretti, V., Bulgarelli, A., Malaguti, G., Spiga, D., & Tiengo, A. (2016). Monte carlo simulations of soft proton flares: testing the physics with xmm-newton. In J.-W. A. den Herder, T. Takahashi, & M. Bautz (Eds.), *Space telescopes and instrumentation 2016: Ultraviolet to gamma ray* (Vol. 9905, p. 99056W). SPIE. doi: 10.1117/12.2232537

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263. Retrieved 2023-10-07, from `http://www.jstor.org/stable/2841583` doi: 10.2307/2841583

Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., & Vasyliunas, V. M. (1994). What is a geomagnetic storm? *Journal of Geophysical Research: Space Physics*, *99*(A4), 5771-5792. doi: 10.1029/93JA02867

Hastie, T., Friedman, J., & Tibshirani, R. (2001). Linear methods for regression. In *The elements of statistical learning: Data mining, inference, and prediction* (pp. 41–78). New York, NY: Springer New York. doi: 10.1007/978-0-387-21606-5_3

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67. Retrieved from `https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634` doi: 10.1080/00401706.1970.10488634

Hubbard, M. W. J., Buggey, T. W., Hall, D., Feldman, C., Keelana, J., Hetherington, O., ... Holland, A. (2024). Techniques for estimating radiation damage from particles passage and focusing from micro pore optics. *Journal of Astronomical Telescopes, Instruments, and Systems*. (In press)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Linear regression. In *An introduction to statistical learning* (1st ed., p. 59-126). Springer New York, NY. doi: 10.1007/978-1-4614-7138-7

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, *110*(A2). doi: 10.1029/2004JA010649

Koga, D., Gonzalez, W. D., Souza, V. M., Cardoso, F. R., Wang, C., & Liu, Z. K. (2019). Dayside magnetopause reconnection: Its dependence on solar wind and

539 magnetosheath conditions. *Journal of Geophysical Research: Space Physics*,
540 *124*(11), 8778-8787. doi: 10.1029/2019JA026889
541 Kronberg, E. A., Clerc, N., Cros, A., de Plaa, J., Gastaldello, F., Gu, L., ... Valen-
542 tini, N. (2020). Prediction and Understanding of Soft-proton Contamination
543 in XMM-Newton: A Machine Learning Approach. *The Astrophysical Journal*,
544 *903*(2), 89. doi: 10.3847/1538-4357/abbb8f
545 Kronberg, E. A., Daly, P. W., Grigorenko, E. E., Smirnov, A. G., Klecker, B., &
546 Malykhin, A. Y. (2021). Energetic charged particles in the terrestrial magneto-
547 sphere: Cluster/rapid results. *Journal of Geophysical Research: Space Physics*,
548 *126*(9), e2021JA029273. doi: 10.1029/2021JA029273
549 Kronberg, E. A., Hannan, T., Huthmacher, J., Münzer, M., Peste, F., Zhou, Z.,
550 ... Ilie, R. (2021). Prediction of soft proton intensities in the near-earth
551 space using machine learning. *The Astrophysical Journal*, *921*(1), 76. doi:
552 10.3847/1538-4357/ac1b30
553 Kronberg, E. A., Kis, A., Klecker, B., Daly, P. W., & Lucek, E. A. (2009). Mul-
554 tipoint observations of ions in the 30–160 kev energy range upstream of the
555 earth's bow shock. *Journal of Geophysical Research: Space Physics*, *114*(A3).
556 doi: 10.1029/2008JA013754
557 Kronberg, E. A., Rashev, M. V., Daly, P. W., Shprits, Y. Y., Turner, D. L., Droz-
558 dov, A., ... Friedel, R. (2016). Contamination in electron observations of the
559 silicon detector on board cluster/rapid/ies instrument in earth's radiation belts
560 and ring current. *Space Weather*, *14*, 449-462. doi: 10.1002/2016SW001369
561 Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statisti-*
562 *cal models* (Fifth ed.). New York: McGraw-Hill Irwin.
563 Luhmann, J. G., Walker, R. J., Russell, C. T., Crooker, N. U., Spreiter, J. R., &
564 Stahara, S. S. (1984). Patterns of potential magnetic field merging sites on
565 the dayside magnetopause. *Journal of Geophysical Research: Space Physics*,
566 *89*(A3), 1739-1742. doi: 10.1029/JA089iA03p01739
567 Milan, S. E., Gosling, J. S., & Hubert, B. (2012). Relationship between interplan-
568 etary parameters and the magnetopause reconnection rate quantified from
569 observations of the expanding polar cap. *Journal of Geophysical Research:*
570 *Space Physics*, *117*(A3). doi: 10.1029/2011JA017082
571 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ...
572 Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of*
573 *Machine Learning Research*, *12*, 2825–2830.
574 Perinati, E., Freyberg, M., Yeung, M. C. H., Pommranz, C., Hess, B., Diebold, S.,
575 ... Santangelo, A. (2024). *Using srg/erosita to estimate soft proton fluxes at*
576 *the athena detectors.*
577 Raab, W., Branduardi-Raymont, G., Wang, C., Dai, L., Donovan, E., Enno, G.,
578 ... Zheng, J. (2016). Smile: a joint esa/cas mission to investigate the inter-
579 action between the solar wind and earth's magnetosphere. In J.-W. A. den
580 Herder, T. Takahashi, & M. Bautz (Eds.), *Space telescopes and instrumen-*
581 *tation 2016: Ultraviolet to gamma ray* (Vol. 9905, p. 990502). SPIE. doi:
582 10.1117/12.2231984
583 Raschka, S., Liu, Y., & Mirjalili, V. (2022). Predicting continuous target variables
584 with regression analysis. In *Machine learning with pytorch and scikit-learn :*
585 *develop machine learning and deep learning models with python* (p. 269-304).
586 Packt Publishing.
587 Read, A. M., & Ponman, T. J. (2003, October). The xmm-newton epic background:
588 Production of background maps and event files. *Astronomy & Astrophysics*,
589 *409*(1), 395–410. doi: 10.1051/0004-6361:20031099
590 Santosa, F., & Symes, W. W. (1986). Linear inversion of band-limited reflection seis-
591 mograms. *SIAM Journal on Scientific and Statistical Computing*, *7*(4), 1307-
592 1330. doi: 10.1137/0907087
593 Shue, J.-H., Chao, J. K., Fu, H. C., Russell, C. T., Song, P., Khurana, K. K., &

Singer, H. J. (1997). A new functional form to study the solar wind control of the magnetopause size and shape. *Journal of Geophysical Research: Space Physics*, *102*(A5), 9497-9511. doi: 10.1029/97JA00196

Sundberg, T., Haynes, C. T., Burgess, D., & Mazelle, C. X. (2016). Ion acceleration at the quasi-parallel bow shock: Decoding the signature of injection. *The Astrophysical Journal*, *820*(1), 21. doi: 10.3847/0004-637X/820/1/21

Tapping, K. F. (2013). The 10.7 cm solar radio flux (f10.7). *Space Weather*, *11*(7), 394-406. doi: 10.1002/swe.20064

Tsyganenko, N. A. (1995). Modeling the earth's magnetospheric magnetic field confined within a realistic magnetopause. *Journal of Geophysical Research: Space Physics*, *100*(A4), 5599-5612. doi: 10.1029/94JA03193

Walsh, B. M., Kuntz, K. D., Collier, M. R., Sibeck, D. G., Snowden, S. L., & Thomas, N. E. (2014). Energetic particle impact on x-ray imaging with xmm-newton. *Space Weather*, *12*(6), 387-394. doi: 10.1002/2014SW001046

Wang, C., Han, J. P., Li, H., Peng, Z., & Richardson, J. D. (2014). Solar wind-magnetosphere energy coupling function fitting: Results from a global mhd simulation. *Journal of Geophysical Research: Space Physics*, *119*(8), 6199-6212. doi: 10.1002/2014JA019834

Wilken, B., Axford, W. I., Daglis, I., Daly, P., Güttler, W., Ip, W. H., ... Ullaland, S. (1997). Rapid. In C. P. Escoubet, C. T. Russell, & R. Schmidt (Eds.), *The cluster and phoenix missions* (pp. 399–473). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-011-5666-0_14

Evaluating Proton Intensities for the SMILE Mission

Simon Mischel[1], Elena A. Kronberg[1], C.P. Escoubet[2]

[1]Department of Earth and Environmental Sciences (Geophysics), Ludwig Maximilian University of Munich (LMU) Munich, Theresienstr. 41, Munich, D-80333, Germany
[2]European Space Research and Technology Centre, Noordwjik, Keplerlaan 1, 2201 AZ, The Netherlands

Key Points:

- Developed models predict proton intensities impacting satellites, aiding space weather forecasting and mission planning.
- Different regions in space showcase distinct relations between proton intensities and predicting parameters.
- Study findings highlight the importance of tailored approaches in space weather prediction.

Corresponding author: Simon Mischel, simonmischel@hotmail.com

Abstract

This study introduces five linear regression models developed to accurately predict proton intensities in the critical energy range of 92.2 keV to 159.7 keV. To achieve this task we utilized 14 years of data sourced from the Cluster's RAPID experiment and NASA's OMNI database. This data was then aligned with the Solar wind-Magnetosphere-Ionosphere Link Explorer (SMILE) mission's trajectory, to increase model accuracy in the relevant regions. Our approach diverges from existing methodologies by offering a user-friendly model that doesn't require specialized software, making it accessible for broader applications in satellite mission planning and risk assessment. The research segregates the dataset into four distinct regions, each analyzed for proton intensity dynamics. In the outer regions ($|\mathrm{YGSE}| \geq 6\,R_e$) there is a pronounced dependence on radial distance and solar wind speed. In contrast, the inner regions ($|\mathrm{YGSE}| \leq 6\,R_e$) demonstrate a significant dependence of proton intensities on the absolute value of the z-coordinate and the magnetic field line topology. Our models achieved a Spearman correlation ranging from 0.57 to 0.72 on the test set, indicating good predictive capabilities. The findings emphasize the role of regional characteristics in space weather prediction and underscore the potential for tailored approaches in future research.

Plain Language Summary

We developed a new model to predict space weather, specifically focusing on proton intensities, which can impact how well satellites work in space. We used 14 years of space observations to create five easy-to-use numerical models. These models are designed to help with planning and protecting future satellite missions, such as the upcoming SMILE mission, from space weather effects. In our study, we looked closely at different areas in space around Earth. We found that in the outer areas ($|\mathrm{YGSE}| \geq 6\,R_e$), the distance from Earth and the speed of the solar wind are important for understanding proton behavior. However, in areas ($|\mathrm{YGSE}| \leq 6\,R_e$), the height above Earth (measured along the z-direction) and the type of magnetic field lines play a more significant role. This shows us that different areas in space around Earth can be affected by space weather in different ways. Our models did a good job of predicting these effects, showing that choosing a tailored approach can be useful when forecasting proton intensities.

1 Introduction

Space weather events, driven by solar activities pose significant challenges to satellite operations and measurements. Notable examples include the European Space Agency's (ESA) Cluster and X-ray Multi-Mirror (XMM-Newton) missions. The Cluster mission, particularly its Research with Adaptive Particle Imaging Detector (RAPID)/Imaging Electron Spectrometer (IES), has encountered challenges due to high proton intensities leading to measurement contamination (Wilken et al., 1997; Kronberg et al., 2016; Kronberg, Daly, et al., 2021). Similarly, the X-ray telescope aboard the XMM-Newton spacecraft experienced significant operational disruptions, with approximately 40% of its observation time compromised due to background contamination (Walsh et al., 2014). Furthermore, investigations into the XMM-Newton telescope's susceptibility to soft protons highlight proton intensities in the sub-100 to 300 keV range, particularly around 100 keV, as the most damaging, leading to significant operational challenges and data contamination, see (Fioretti et al., 2016) and references therein.

The upcoming European-Chinese Solar wind-Magnetosphere-Ionosphere Link Explorer (SMILE) mission (Branduardi-Raymont et al., 2018), slated for launch in 2025, aspires to deepen our understanding of the Sun-Earth interaction, decoding space weather hazards and understanding energy entry into Earth's magnetosphere. While missions like ATHENA (Advanced Telescope for High Energy Astrophysics), which will maintain an orbit around the L2 Lagrange point with a continuously large distance to Earth, are ex-

pected to face minimal threats from soft protons (Perinati et al., 2024), SMILE with its highly inclined elliptical orbit around earth, akin to that of Cluster and XMM-Newton, will navigate through diverse magnetospheric regions, making it susceptible to soft proton radiation. Particularly its Soft X-ray Imager (SXI) telescope, presents challenges concerning radiation exposure (Raab et al., 2016; Branduardi-Raymont & Wang, 2022). Therefore a critical component of achieving SMILE's objectives is the accurate prediction of proton radiation levels, which significantly affect the Total Ionizing Dose (TID) and Total Non-Ionizing Dose (TNID) absorbed by the Charge-Coupled Devices (CCDs) of the Soft X-ray Imager (SXI) (Hubbard et al., 2024). In recent discussions (M. Hubbard, personal communication, 2023) the necessity for models that accurately estimate radiation levels was emphasized, particularly in critical energy ranges below 300 keV, crucial for the SMILE mission's success. A strong preference for models that are not only accurate but also straightforward and interpretable was expressed.

To meet these needs, our study adopts a distinct approach compared to existing research, which is based on machine learning black box models, such as the works of Kronberg et al. (2020) and Kronberg, Hannan, et al. (2021). We aim to develop a simple, user-friendly linear regression model leveraging data from the Cluster mission and NASA's OMNI database (King & Papitashvili, 2005). Our model's simplicity and ease of use make it accessible to a broader range of users, not requiring specialized software or extensive computational resources. This approach not only contributes to the scientific understanding of space weather phenomena but also offers practical tools for satellite mission planning and risk assessment.

2 Data Analysis and Processing

2.1 Data Preparation: Adapting CLUSTER's Dataset for the SMILE Mission's Trajectory

The proton intensity data for this research was taken from the Cluster's RAPID experiment, ranging from 2001 to 2015. The experiment captures 3-D energetic electron and ion fluxes above approximately 30 keV using the Imaging Electron Spectrometer (IES) and the Imaging Ion Mass Spectrometer (IIMS) instruments. Situated in the SCENIC detector head, the IIMS instrument identifies ion energies and species. The methodology involves using start and stop signals produced from electrons emitted by an initial thin foil on the solid-state detector's surface. The time-of-flight (TOF) between these signals, combined with the known energy, discerns the species and energy channel (Daly & Kronberg, 2023). This study specifically uses data from the *p3* channel, targeting proton intensities between 92.2 keV and 159.7 keV. As predicting parameters for solar, solar wind and geomagnetic activity we used variables from the OMNI database(`https://omniweb.gsfc.nasa.gov/`), see also King and Papitashvili (2005).

Aiming for a model tailored to the SMILE mission's trajectory (see Figure 1), data filtering was imperative. Points not adhering to the following spatial parameters were excluded: $-10.5R_e \leq x \leq 11.2R_e$; $-10.8R_e \leq y \leq 11.5R_e$; $z \leq 18.5R_e$; $\sqrt{x^2+y^2} \leq 11.6R_e$; $\sqrt{x^2+z^2} \leq 19.8R_e$; $\sqrt{z^2+y^2} \leq 20.0R_e$. These constraints were chosen by rounding the maxima and minima of the spatial parameters to the nearest tenth. As a result, we were left with a trimmed dataset, reduced from 1,172,923 to 462,615 data points. Though compact, this dataset is centered on the space region significant for SMILE, promising heightened model accuracy. It's noteworthy that negative z-values weren't excluded, considering the SMILE mission's highly inclined, elliptical orbit, which dips to -3.5 R_e. Omitting these would disregard vital data, especially since the Cluster's trajectory spent a notable amount of time in the southern hemisphere.

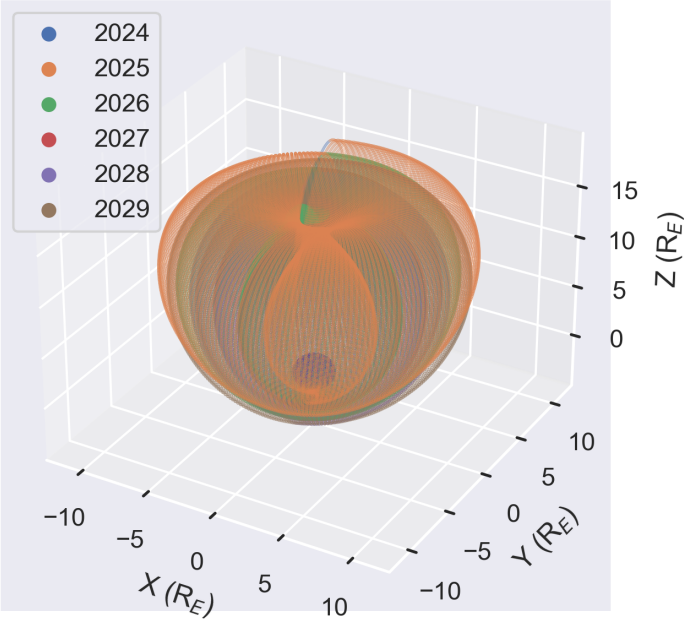


Figure 1. SMILE mission's trajectory. Distinct colors represent individual years with Earth, having a radius of 6,731 km, at the center.

2.2 Predictor Introduction

The spacecraft's location in the Geocentric Solar Ecliptic (GSE) coordinate system is defined by `x`, `y`, and `z` in Earth radii (R_e). The variable `rdist` denotes the satellite's radial distance from Earth. The magnetic field line type, termed as "Foot Type", indicates the connectivity of the IMF field lines to Earth, calculated using the Tsyganenko (1995) model. The initial definition stated by Kronberg et al. (2020) is as follows: the interplanetary magnetic field lines (IMF) with no connection to Earth have Foot Type 0, open magnetic field lines with one connection to Earth have Foot Type 1, and closed field lines with both ends connected to Earth have Foot Type 2. It was, however, decided to redefine the IMF to 1 and open field lines to 0, to achieve a stronger linear relationship between Foot Type and the target variable, as discussed in chapter 2.3.2.

The Disturbance storm time index (`Dst_index`) characterizes geomagnetic storms in the unit nT (Banerjee et al., 2012). The Auroral Electrojet (`AE_index`) quantifies magnetic activity in the auroral zone, also denoted in nT. The 10.7 cm solar radio flux (F10.7) with unit sfu serves as a solar activity level indicator and a proxy for solar emissions (Tapping, 2013).

The IMF direction is described by its components `BimfxGSE`, `BimfyGSE`, and `BimfzGSE` in the GSE system in nT. The IMF direction at the magnetopause determines if reconnection happens on the dayside (Crooker et al., 1979; Luhmann et al., 1984; Koga et al., 2019). Plasma properties of the solar wind are described by Solar wind speed (`VSW`) in km/s, proton density (`NpSW`) in cm^{-3}, and temperature (`Temp`) in K. The direction of the solar wind velocity is described by `VxSW_GSE`, `VySW_GSE`, and `VzSW_GSE`. The solar wind dynamic pressure, `Pdyn` (nPa), can be represented as:

$$\texttt{Pdyn} = \texttt{NpSW} * \texttt{VSW}^2 * 1.67 * 10^{-6} \quad (1)$$

2.3 Exploratory Data Analysis

2.3.1 Spatial Proton Intensity Distribution

To analyze the proton intensity distribution in relation to the spacecraft's trajectory, we combined the y and z coordinates to produce a radial distance variable, termed `yz_axis`. We introduced this variable because Cluster's trajectory is predominantly in the southern hemisphere, contrasted with SMILE's expected northern trajectory. The `yz_axis` is computed by $\sqrt{y^2 + z^2}$, offering a simplified yet informative perspective on proton intensity's spatial distribution. Figure 2 depicts the spatial distribution of proton intensities in the x, $\sqrt{y^2 + z^2}$ coordinate system. The color gradient represents the percentage of measurements that exceed 2, the mean value of $\log_{10}$(proton intensities) rounded to one significant digit, highlighting regions with prolonged high proton intensities. The central black void indicates missing measurements. This gap arises from our deliberate exclusion of data points with radial distances (`rdist`) below 6 Earth radii (R_e) in order to emphasize regions beyond the radiation belts. Historically, proton intensities surge in zones under 6 R_e, which encompass the ring current and radiation belt regions. Our focus shifts to lesser-studied areas, with their generally lower intensities outside of the radiation belts, mainly because the SXI telescope on the SMILE mission is equipped with a shutter mechanism, protecting its Charge Coupled Devices (CCD) from intense radiation within the radiation belt. The decline in proton intensities with increased distance from Earth is observed, irrespective of whether it's along the `x` axis or the y-z plane. This observation aligns with subsequent feature plots and correlation matrix analyses. Moreover, this analysis reveals areas along closed magnetic field lines with heightened proton intensities, as well as sparser regions corresponding to open magnetic field lines over the polar cap, demonstrating a clear spatial correlation between magnetic field line configuration and proton intensity distribution.

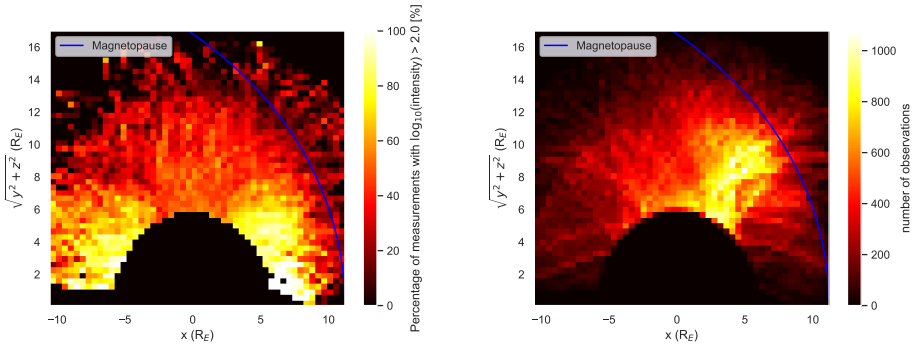


Figure 2. (Left) Heatmap of proton intensities against the x-coordinate and $\sqrt{y^2 + z^2}$. The color gradient represents the percentage of measurements where $\log_{10}$(proton intensity) > 2, which is the mean value rounded to one significant digit. (Right) Data point density for each bin, with the blue line representing the magnetopause, derived using Shue et al. (1997). Both plots incorporate the 462,615 data points post-reshaping for the SMILE mission.

2.3.2 Cross-Correlation and Feature Plot Analysis

Cross-correlation matrices, employing the Pearson coefficient, are used in feature selection for linear regression models. The coefficient quantifies the linear relationship

strength and direction between two variables, spanning from -1 (perfect negative relationship) to 1 (perfect positive relationship), with 0 indicating no linear correlation. Such analyses illuminate potential multicollinearity issues in datasets, which can adversely affect regression coefficient stability and model interpretability (Raschka et al., 2022; James et al., 2013).

Analyzing the cross-correlation matrix (Figure 3), we observed:

- `FootType`: The feature plot in Figure 4 highlighted a clear potential for refining the correlation between `FootType` and `p3`. The initial positive correlation of 0.24 with `p3` was improved to 0.41 upon redefining the foot type as mentioned in section 2.2.
- `AE_index`: While the correlation was weak (0.07), the feature plot identified AE values surpassing 2600 nT as possible outliers.
- F10.7 solar radio flux index: Given its correlation coefficient of -0.20, the feature plot shows a predominantly monotonically decreasing relationship between F10.7 and the target variable.
- `VxSW_GSE`: The feature plot demonstrated that proton intensity increases with higher absolute wind speeds up to 950 km/s. Values exceeding this were considered as potential outliers.
- Distance variables: While `z` showed a positive correlation of 0.21, its relationship with proton intensities displayed a clear maximum around 0 on the feature plot. This insight led to the introduction of `|z|` as an improved predictor.
- Other Variables: The strong negative correlations of `rdist` and `yz_axis` with `p3` were supported by the feature plot's linear regression lines, emphasizing their importance as predictors.

Analysis of the proton intensity histogram identified two extreme outliers exceeding 100,000 $1/cm^2/s/sr/keV$ (see Figure A1). Removing these and other above-identified outliers from different predictors did not improve model performance, justifying their retention. Further analysis revealed 606 F10.7 measurements above 900, deemed unrealistic and consequently removed.

2.4 Data Split and Data Scaling

This section outlines the additional processing steps applied to the data set, already reshaped for the SMILE mission as detailed in section 2.1. These steps include splitting the data into training and testing sets, transforming the target variable, and scaling features.

Records before December 31, 2012, were allocated to the training set, while records from January 1, 2013, onwards formed the test set. This temporal division results in an approximate 75% to 25% split between the training and test datasets. The training set, was then later further divided into training and validation sets by the use of five-fold cross-validation, where the dataset is divided into five parts, with each part being used as a validation set while the remaining four parts are used as training data

The proton intensities recorded in channel 3 (`p3`), our target values, display a wide spectrum. We therefore transformed these values using a base 10 logarithmic function. Addressing the challenge of logging zero values, all such occurrences in `p3` were substituted with 0.5. However, this introduces potential pitfalls as we expect an artificial population with the same values, a concern later revisited during model evaluation (Bellégo et al., 2021).

Optimizing gradient descent requires careful attention to feature scaling (Raschka et al., 2022). In our polynomial regression model, we employed a double-scaling tech-

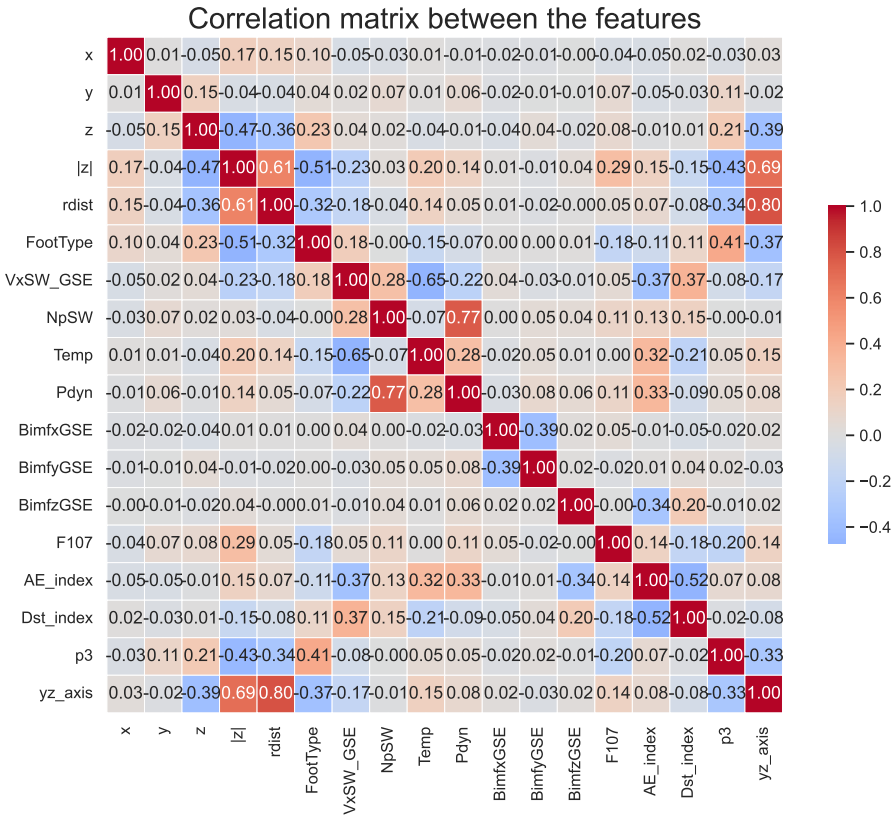


Figure 3. Pearson coefficient-based correlation matrix for the predictors and the proton intensity post-data reshaping for the SMILE mission.

nique to ensure numerical stability and facilitate the optimization process. Initially, the original features were scaled to zero mean and unit variance using the `StandardScaler()` method, aligning with the desired outcomes (Pedregosa et al., 2011). Subsequently, polynomial features were generated from these scaled features. To further enhance the model's robustness, these polynomial features were subjected to a second round of scaling using the same `StandardScaler()` method.

By scaling both the original and polynomial features, we ensure that the coefficients are directly comparable in terms of their contribution to the model and that all features display a mean and unit variance of zero.

3 Methodology

3.1 Linear Regression Model

The choice of employing linear regression models in this study is underpinned by several reasons. First and foremost, linear regression models offer a simple and interpretable framework for understanding how input variables affect the output. Furthermore, the methodology allows for the transformation of input variables to enhance their predictive capabilities, such as the introduction of polynomial terms and interaction effects (Hastie et al., 2001).

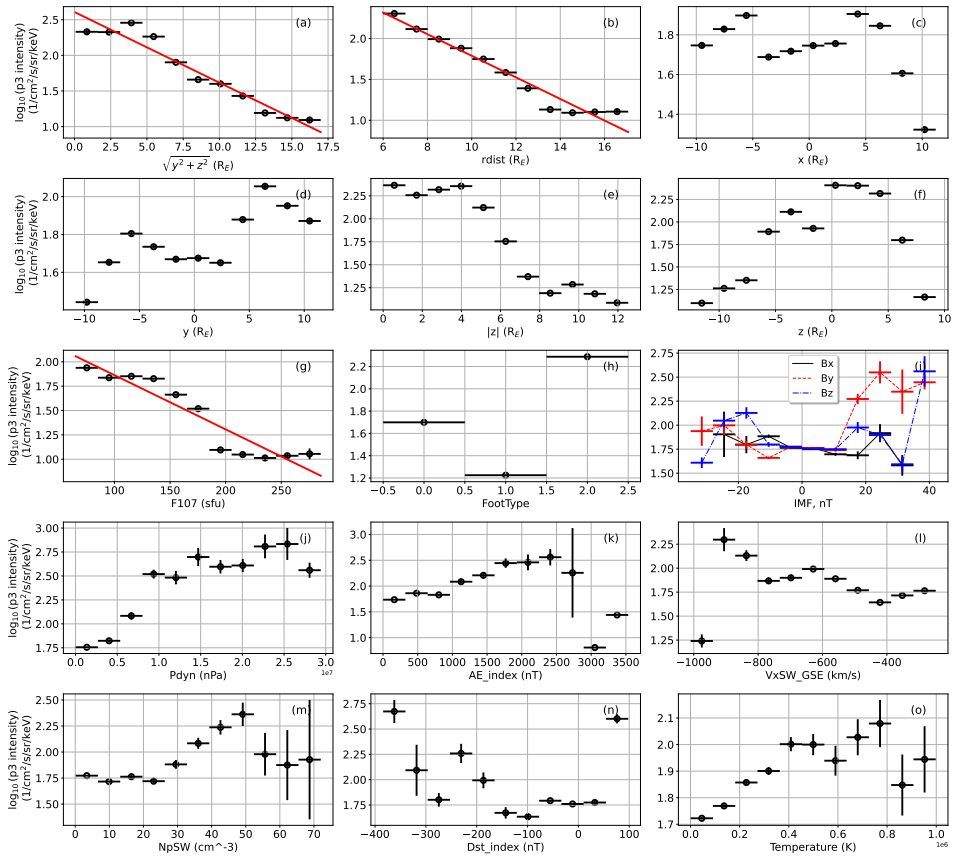


Figure 4. Mean of the logarithmically scaled proton intensities from the `p3` channel against potential predictors. Vertical lines depict the standard 95% confidence level, while horizontal lines indicate bin half-widths. Linear regression lines in red are shown for `rdist`, `yz_axis`, and F10.7.

The Ordinary Least Squares (OLS) model serves as the foundational approach, focusing on minimizing the sum of squared differences between observed and predicted values (James et al., 2013; Galton, 1886). To tackle the possible issue of multicollinearity, Ridge Regression can be utilized, which incorporates an L2 penalty term into the loss function (Kutner et al., 2005; Hoerl & Kennard, 1970). Lasso Regression is employed when feature selection is essential, as it uses an L1 penalty to drive certain coefficients to zero, effectively eliminating them from the model (Santosa & Symes, 1986). Lastly, Elastic Net Regression can be used to combine the strengths of both L1 and L2 penalties, providing a balanced approach that can handle both multicollinearity and feature selection (Pedregosa et al., 2011). Multiple models were trained using the scikit-learn library in Python (Pedregosa et al., 2011).

3.2 Model Selection and Optimization

For model evaluation, we utilized a set of metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2 (coefficient of determination), Pearson correlation, and Spearman correlation. Model selection was primarily guided by the performance of R^2 and Spearman correlation on the validation set. To ensure a robust and generalizable evaluation, five-fold cross-validation with the help of the `KFold` function from `sklearn.model_selection` was applied to the training set. Given the time-series nature of our dataset, the `shuffle` parameter within the cross-validation procedure was intentionally set to `false`. Subsequently, the evaluation metrics were computed as the average values derived from the five cross-validation folds, thereby offering a more reliable measure of the model's true performance.

3.2.1 Simple OLS, Lasso, Ridge and Elastic Net

Following the initial selection of linear regression models, two distinct approaches were undertaken to optimize model performance. The first approach involved the application of various linear regression techniques, including OLS, Lasso, Ridge, and Elastic Net. This approach, however, did not yield satisfactory results. The maximum R^2 value on the validation set was only 0.02, and the highest Spearman correlation coefficient was 0.43.

3.2.2 Introduction of Polynomial Terms

To improve upon this, the second approach incorporated polynomial terms into a standard Lasso model from `sklearn.linear`. The model was optimized for the regularization parameter α using five-fold cross-validation. The optimal α was determined using `LassoCV` with a maximum iteration of 10,000 and a tolerance of 1×10^{-5}. This approach significantly improved the model performance, achieving an R^2 value of 0.22 and a Spearman correlation coefficient of 0.51 on the validation set.

3.2.3 Heuristic-based Feature Selection Technique

However, this model included 52 predictors, making it complex and potentially prone to overfitting. Further work was needed to develop a more parsimonious model with a maximum of 25 predictors while maintaining acceptable performance. To reduce the number of predictors while maintaining model performance, we adopted a heuristic-based feature selection strategy. For this strategy the Stochastic Gradient Descent (SGD) framework was employed, with the algorithm configured as follows: the regularization term (α) was set to the optimal value identified through cross-validation. The learning rate was set to a constant value, initialized at $\eta_0 = 1 \times 10^{-5}$. The hyperparameter defining the loss function was set to the squared error loss. An L1 penalty term was incor-

porated for feature selection. The algorithm was set to terminate when the tolerance reached 1×10^{-5}, with a maximum of 100 iterations for convergence.

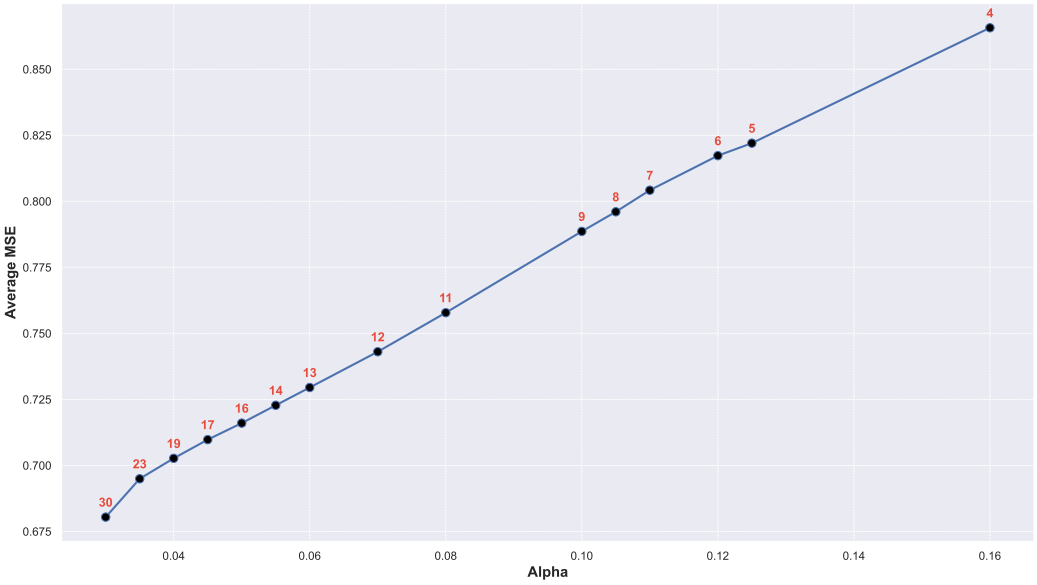


Figure 5. Plot of Average Mean Squared Error (MSE) against the regularization parameter α. The curve exhibits an "elbow" point at 13 predictors, indicating a minimal but acceptable loss in model performance. A noticeable increase in MSE is observed when the number of predictors is reduced from 13 to 12, suggesting that all 13 predictors left, display significant importance for the model. This "elbow" point serves as the basis for selecting an optimal α value and, consequently, the number of predictors for the final model.

Unlike earlier approaches that solely aimed to minimize the Mean Squared Error (MSE), this method also considers the number of predictors in the final model. We tested a range of regularization parameters (α) and sought to identify a "knee" or "elbow" in the plot of MSE versus α. This point represents a compromise between model performance and complexity.

To enhance the robustness of the feature selection process, we employed K-Fold cross-validation with the shuffle parameter set to `True`. This approach allows for a more representative sampling of the training data across each fold. Specifically, we aimed to identify the most stable set of predictors corresponding to the "elbow" point for the regularization parameter α. By enabling shuffling during cross-validation, we increase the likelihood that the predictor set extracted from one of the folds offers a more comprehensive representation of the entire training dataset. The α range chosen was from 0.03 to 0.17, which covered all models with the amount of non-zero predictors ranging from 28 to 4.

Upon employing this approach, we identified a subset of 13 predictors by analyzing the MSE vs α plot in Figure 5. Importantly, we operate under the assumption that all predictors remaining after the feature selection process are relevant to the outcome. Therefore, penalizing these predictors, as Lasso does, could introduce an unwanted bias into the model. Given this consideration, an OLS model was chosen for the final training rather than a Lasso regression model.

Table 1. Average performance metrics for different models resulting from five-fold cross-validation and the number of data points N_train used for training.

Model	MSE	MAE	R^2	Pearson	Spearman	Predictors	N_train
Basic_OLS	0.92	0.76	0.02	0.43	0.43	9	353,660
Poly_Lasso	0.73	0.68	0.22	0.51	0.51	52	353,660
Heuristic_Poly_OLS	0.78	0.71	0.17	0.47	0.47	13	353,660
Split_Poly_Part1	0.58	0.60	0.24	0.53	0.53	5	60,961
Split_Poly_Part2	0.82	0.73	0.09	0.50	0.50	19	145,952
Split_Poly_Part3	0.79	0.72	0.19	0.46	0.46	15	70,137
Split_Poly_Part4	0.65	0.64	0.29	0.55	0.55	6	76,610

Although the resulting model exhibits lower performance on the validation set, as evidenced by Table 1, it better aligns with the study's objectives of interpretability and usability compared to the Lasso model with 52 predictors. This heuristic-based feature selection strategy aligns well with the principle of Occam's razor, suggesting that simpler models are preferable when performance is comparable. Therefore, this approach effectively strikes a balance between the number of predictors and model performance, thereby enhancing the model's interpretability and practical utility.

3.2.4 Data Split

An in-depth analysis of the relationship between the `y` and `p3` variables revealed that the data could be divided into four distinct parts, each characterized by an increasing or decreasing slope, see Figure 4 (d). This lead to the decision to split the dataset into four separate parts based on specific conditions, as described below:

- Part 1: $\mathtt{y} \leq$ -6.6 R_e
- Part 2: -6.6 $R_e \leq \mathtt{y} \leq$ 2.3 R_e
- Part 3: 2.3 $R_e \leq \mathtt{y} \leq$ 6 R_e
- Part 4: $\mathtt{y} \geq$ 6 R_e

Upon splitting the data, separate models were built for each part, using the same heuristic-based predictor selection technique previously described. These outperformed the non-split OLS model in the Spearman correlation coefficient and the R^2 metric for three out of the four subsets (see Table 1), all while maintaining low model complexity.

4 Results

This chapter presents the empirical results obtained from the evaluation of various OLS models on the unseen test set. The models are compared based on a set of evaluation metrics and feature importances.

4.1 Presentation of Final Models

In this section, we present the final forms of our linear regression models developed for predicting proton intensities. Each model is displayed with its coefficients in basic, unscaled units, offering a clear view of the relative impact of each predictor variable. These models encapsulate our findings and are ready for practical application.

Heuristic_Poly_OLS:

$$\begin{aligned}\log_{10}(p3) = {} & 2.1 \times 10^{-2} \cdot y - 3.0 \times 10^{-1} \cdot |z| - 1.7 \times 10^{-1} \cdot \text{rdist} + 1.1 \cdot \text{FootType} \\ & - 2.0 \times 10^{-3} \cdot \text{VxSW\_GSE} + 6.2 \times 10^{-8} \cdot \text{Pdyn} - 1.1 \times 10^{-3} \cdot \text{F107} \\ & + 3.7 \times 10^{-4} \cdot \text{AE\_index} + 8.3 \times 10^{-4} \cdot x \cdot |z| + 1.8 \times 10^{-2} \cdot |z| \cdot \text{rdist} \\ & - 7.4 \times 10^{-2} \cdot \text{rdist} \cdot \text{FootType} + 7.7 \times 10^{-4} \cdot \text{rdist} \cdot \text{F107} \\ & - 3.1 \times 10^{-5} \cdot \text{F107}^2 + 2.3 \end{aligned} \tag{2}$$

Split_Poly_Part1:

$$\begin{aligned}\log_{10}(p3) = {} & -5.9 \times 10^{-1} \cdot \text{rdist} + 8.4 \times 10^{-1} \cdot \text{FootType} - 2.5 \times 10^{-3} \cdot \text{VxSW\_GSE} \\ & + 2.0 \times 10^{-2} \cdot \text{rdist}^2 - 5.0 \times 10^{-2} \cdot \text{rdist} \cdot \text{FootType} + 4.2 \end{aligned} \tag{3}$$

Split_Poly_Part2:

$$\begin{aligned}\log_{10}(p3) = {} & -6.2 \times 10^{-1} \cdot |z| + 2.3 \times 10^{-1} \cdot \text{rdist} + 7.5 \times 10^{-1} \cdot \text{FootType} \\ & + 7.8 \times 10^{-4} \cdot \text{VxSW\_GSE} + 1.3 \times 10^{-7} \cdot \text{Pdyn} + 3.0 \times 10^{-4} \cdot \text{AE\_index} \\ & + 3.4 \times 10^{-3} \cdot x \cdot |z| + 3.4 \times 10^{-5} \cdot x \cdot \text{VxSW\_GSE} - 5.3 \times 10^{-4} \cdot y \cdot \text{rdist} \\ & + 5.1 \times 10^{-2} \cdot |z| \cdot \text{rdist} - 3.7 \times 10^{-2} \cdot \text{rdist}^2 - 7.2 \times 10^{-2} \cdot \text{rdist} \cdot \text{FootType} \\ & - 2.5 \times 10^{-4} \cdot \text{rdist} \cdot \text{VxSW\_GSE} + 6.0 \times 10^{-4} \cdot \text{rdist} \cdot \text{F107} \\ & + 1.8 \times 10^{-1} \cdot \text{FootType}^2 - 5.7 \times 10^{-8} \cdot \text{FootType} \cdot \text{Pdyn} \\ & + 9.2 \times 10^{-14} \cdot \text{Temp} \cdot \text{Pdyn} \quad - 1.9 \times 10^{-16} \cdot \text{Pdyn}^2 - 2.3 \times 10^{-5} \cdot \text{F107}^2 + 1.9 \end{aligned} \tag{4}$$

Split_Poly_Part3:

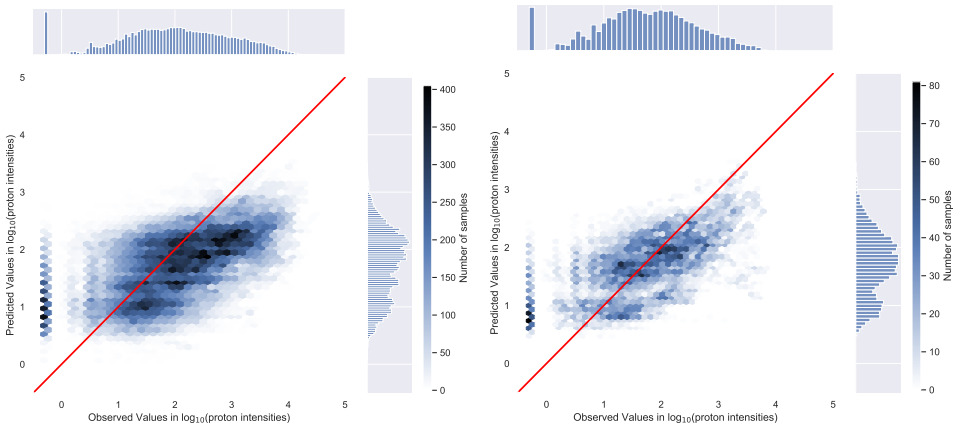
$$\begin{aligned}\log_{10}(p3) = {} & 3.6 \times 10^{-1} \cdot y - 2.8 \times 10^{-1} \cdot |z| - 5.7 \times 10^{-2} \cdot \text{rdist} + 1.1 \cdot \text{FootType} \\ & - 1.7 \times 10^{-3} \cdot \text{VxSW\_GSE} + 2.5 \times 10^{-7} \cdot \text{Temp} + 5.2 \times 10^{-8} \cdot \text{Pdyn} \\ & - 1.3 \times 10^{-3} \cdot \text{F107} + 4.6 \times 10^{-4} \cdot \text{AE\_index} + 2.5 \times 10^{-3} \cdot x \cdot |z| \\ & - 2.8 \times 10^{-2} \cdot y \cdot \text{rdist} + 1.4 \times 10^{-2} \cdot |z| \cdot \text{rdist} - 8.2 \times 10^{-2} \cdot \text{rdist} \cdot \text{FootType} \\ & + 8.2 \times 10^{-4} \cdot \text{rdist} \cdot \text{F107} - 3.1 \times 10^{-5} \cdot \text{F107}^2 + 1.2 \end{aligned} \tag{5}$$

Split_Poly_Part4:

$$\begin{aligned}\log_{10}(p3) = {} & -5.1 \times 10^{-2} \cdot |z| - 1.6 \times 10^{-1} \cdot \text{rdist} + 2.2 \times 10^{-1} \cdot \text{FootType} \\ & - 2.6 \times 10^{-3} \cdot \text{VxSW\_GSE} - 1.9 \times 10^{-3} \cdot \text{F107} + 7.4 \times 10^{-4} \cdot \text{AE\_index} + 2.7 \end{aligned} \tag{6}$$

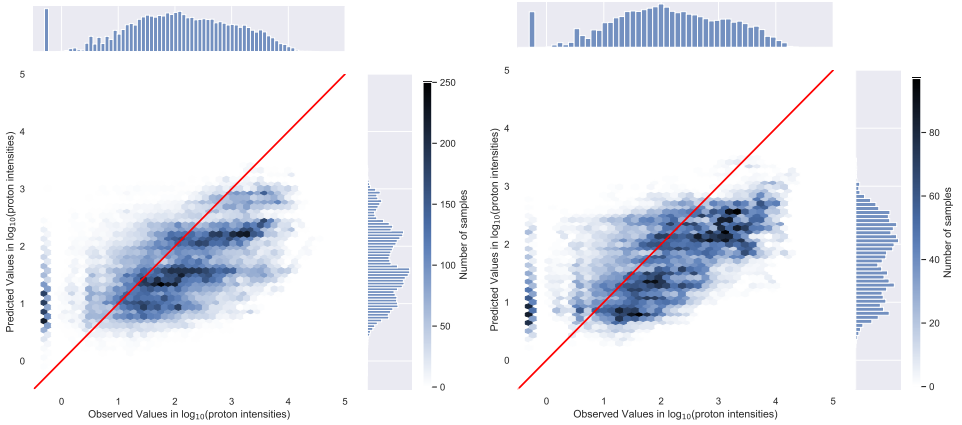
4.2 Performance on the Test Set

In the Heuristic_Poly_OLS model, the hexagonal bins largely align with the ideal fit line (see Figure 6), which is indicative of good predictive performance. However, this model exhibits a tendency to underestimate observed values, notably at higher proton intensities. A significant peak at log(0.5) in the histogram of observed values is associated with an overestimation in the Heuristic_Poly_OLS model's predictions. This peak stems from the substitution of zero values in the target variable $p3$ before applying the logarithmic transformation. This overestimation at log(0.5) potentially skews the model's learning process, causing it to adjust its predictions downward to minimize the overall loss. While this adjustment mitigates the error for overestimated values, it concurrently introduces a bias leading to the underestimation of other observed values. This behav-



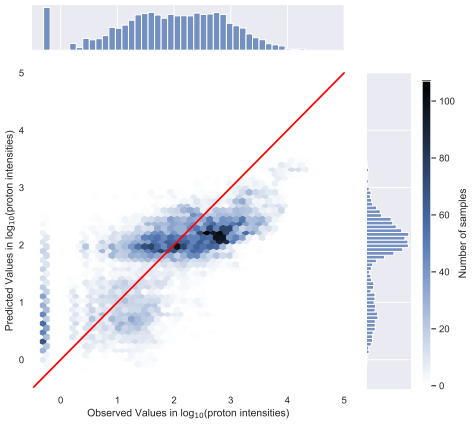
(a) Heuristic_Poly_OLS model

(b) Split_Poly_Part1



(c) Split_Poly_Part2

(d) Split_Poly_Part3



(e) Split_Poly_Part4

Figure 6. Jointplots comparing observed and predicted values of proton intensities, of the test set for the different OLS models. The red lines represent ideal fits where observed values equal predicted values. Color bars indicate the number of samples in each hexagonal bin. Histograms at the top and right margins show the distributions of observed and predicted values for each model.

Table 2. Performance metrics of the final models on test data.

Model	MSE	MAE	R^2	Pearson	Spearman	Predictors
Heuristic_Poly_OLS	0.74	0.71	0.22	0.56	0.57	13
Split_Poly_Part1	0.46	0.54	0.38	0.62	0.61	5
Split_Poly_Part2	0.83	0.74	0.11	0.56	0.57	19
Split_Poly_Part3	0.77	0.73	0.24	0.61	0.62	15
Split_Poly_Part4	0.47	0.56	0.50	0.72	0.72	6

ior is consistently observable across all models, particularly those focusing on specific regions. As an alternative to zero substitution, we also explored the removal of these zero values. While this approach enhanced performance on the training set, it consistently led to diminished performance on the validation set. Consequently, despite its limitations, the zero substitution technique was retained to ensure better generalization to unseen data.

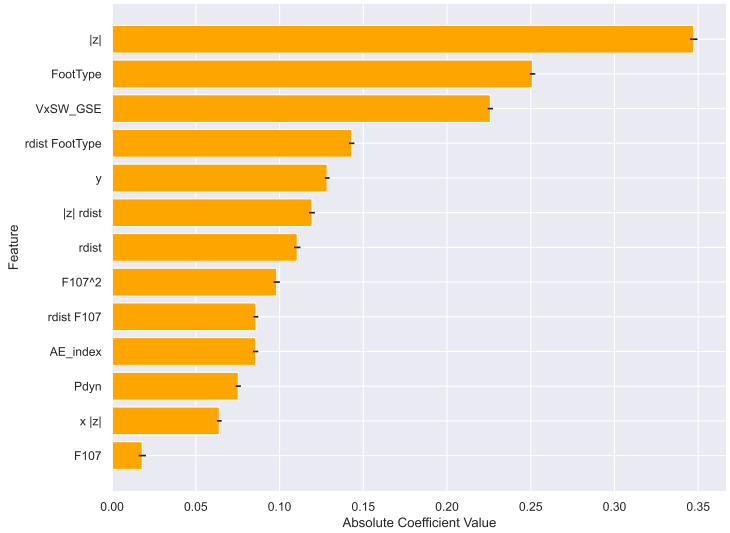
Turning our attention to Split_Poly_Part1 and Split_Poly_Part4, these models exhibit the most well-centered distribution around the ideal fit line in their respective heatmaps. This observation is consistent with their performance metrics as recorded in Table 2, showcasing R^2 values of 0.38 and 0.50 and Spearman coefficients of 0.61 and 0.72 for the test set. Notably, these models also maintain low complexity, employing only 5 and 6 predictors, respectively.

Conversely, Split_Poly_Part2, with its high complexity due to having 19 predictors, exhibits subpar performance despite an acceptable Spearman coefficient. Significantly, with an R^2 value of only 0.11, this model is the sole split variant that exhibits notably inferior performance compared to the unsplit Heuristic_Poly_OLS model in the test set. The more inhomogeneous distribution of hexagonal bins in its heatmap is indicative of this weaker performance. On the other hand, Split_Poly_Part3 shows a modest improvement over the unsplit model. This is evident not only in the performance metrics but also in a more concentrated distribution in its heatmap, compared to Split_Poly_Part2.

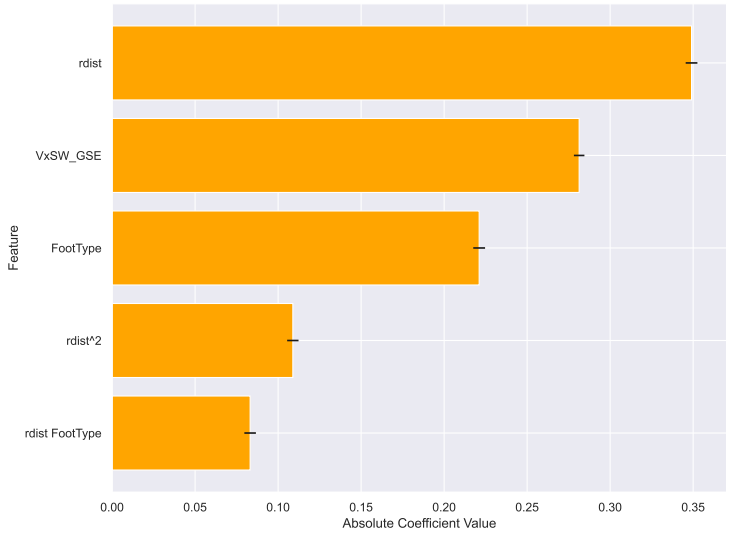
4.3 Feature Importance

In order to derive feature importance in a linear regression model, one can examine the coefficients of the model. The magnitude of the coefficients indicates the relative importance of the corresponding feature in predicting the target variable. A larger absolute value of a coefficient suggests a stronger influence of the associated feature on the outcome. The features are scaled appropriately by scaling the features once before the creation of the polynomials and once afterward. Scaling ensures that all features are on a comparable scale, which prevents features with larger values from dominating those with smaller values in the model. The feature importance for each model was plotted in figure 7.

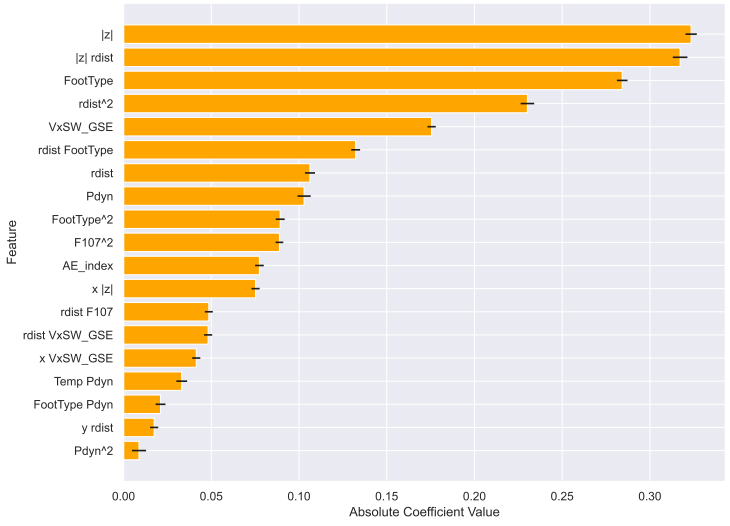
The variations in feature importance across the different models offer insights into the underlying mechanisms affecting proton intensities in various regions. For the Heuristic_Poly_OLS model and models corresponding to the inner regions (Split_Poly_-Part2 and Split_Poly_Part3), the absolute value of z (`|z|`) emerges as the most significant predictor, next to `Foottype` and `VxSW_GSE`. The Split_Poly_-Part2 model additionally identifies the polynomial terms `|z| rdist` and $\texttt{rdist}^2$ as significant features. The similarity between the models for the inner part and the model trained on the full data is most likely partially influenced by the fact that the inner regions contain 61% of the total data points. The models tailored to the outer regions (Split_Poly_Part1 and Split_Poly_Part4)



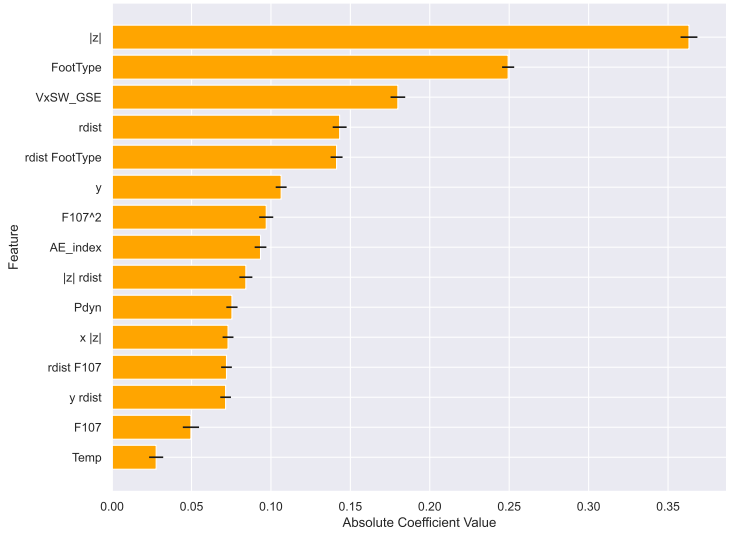
(a) Heuristic_Poly_OLS model



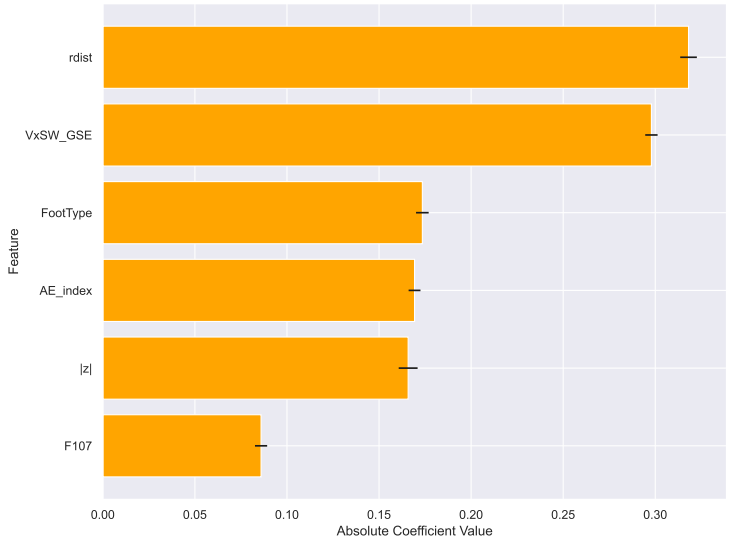
(b) Split_Poly_Part1



(c) Split_Poly_Part2



(d) Split_Poly_Part3



(e) Split_Poly_Part4

Figure 7. Feature importance plots for five different OLS models: Each plot presents the absolute values of the model coefficients, serving as indicators of feature importance. Accompanying error bars represent the standard errors, providing a measure of the coefficient's reliability. The plots collectively offer insights into the relative significance of each predictor across different models.

prioritize `rdist`, `VxSW_GSE`, and `Foottype` as their top predictors, in that order. A notable distinction from the inner region models is the elevated significance of `VxSW_GSE`.

5 Discussion

The most critical predictor for our Heuristic_Poly_OLS_Model, which utilizes the full dataset, is the absolute value of z, denoted as $|z|$. The model reveals a negative correlation between $|z|$ and the proton intensities, indicating that as $|z|$ increases, the proton intensity declines. This trend can be primarily attributed to the circulation of protons in Earth's magnetic field. Most ions are concentrated at the equatorial plane during their drift trajectories on the closed magnetic field lines. At higher latitudes where open magnetic field lines dominate, the proton intensities are expected to drop with $|z|$ distance. Consequently, the proton intensities reduce with an increase in $|z|$.

The predictor `FootType` categorizes magnetic field line types and ranks as the second most influential factor. Closed field lines, known for the highest proton intensities, trap charged particle populations. The importance of this parameter aligns with the studies by Walsh et al. (2014) and Kronberg et al. (2020). In contrast, open field line regions typically correlate with lower particle energies outside the soft proton (SP) range, resulting in weaker count rates as detailed in (Kronberg et al., 2020). IMF regions show slightly higher count rates since particles can experience acceleration in the bow shock region, especially quasi-parallel bow shock configurations (normal to the shock is parallel to the IMF direction) (Blandford & Ostriker, 1978; Kronberg et al., 2009; Sundberg et al., 2016).

The high importance of solar wind speed in the X-direction is consistent with the analysis of the feature plot in figure 4. Kronberg, Hannan, et al. (2021) also found that `VxSW_GSE` displays "the most substantial linear dependence of the proton intensities among the OMNI parameters." The solar wind speed, directly correlated to its electric field as $E = V_x \times B_z$, is crucial for magnetospheric dynamics as it determines the rate of magnetic reconnection on Earth's dayside (Dorelli, 2019), and consequently magnetic reconnection at the night side. A surge in solar wind speed correlates with an increased rate of magnetic reconnection. Additionally, magnetic reconnection events, which can accelerate charged particles, also impact soft proton intensities significantly, as noted by (Read & Ponman, 2003). Research by Gonzalez et al. (1994), Milan et al. (2012)), and Wang et al. (2014) further elucidates this concept, indicating that a variety of solar wind-magnetosphere energy transfer models are dependent on the velocity of the solar wind.

6 Conclusion

In this study, we developed five user-friendly linear regression models to predict proton intensities in the energy range of 92.2 keV to 159.7 keV with a Spearmen correlation ranging from 0.57 to 0.72 on the test data. Utilizing data from the Cluster's RAPID experiment, supplemented with solar, solar wind, and geomagnetic data from the OMNI database, the study focused on aligning the models with the anticipated spatial area covered by the upcoming SMILE-mission.

Segmenting the data into four distinct regions based on the y coordinate with thresholds -6.6 R_E, 2.3 R_E and 6 R_E, resulted in enhanced model performance for three of the four segments, surpassing the main model's performance. The primary predictors in these outer regions were identified as radial distance and the radial solar wind speed. Conversely, the inner region models and the comprehensive main model demonstrated a significant dependence on the absolute value of z and the type of magnetic field lines.

The redefinition of the `FootType` variable and the incorporation of the absolute value of z as key model features significantly improved the model compared to previous relevant studies. This study suggests that the development of more accurate predictive mod-

els for space weather phenomena may not solely rely on novel algorithms, but also on crafting tailored models, each addressing distinct regions with their specific characteristics.

Appendix A : Histogram of Proton Intensities

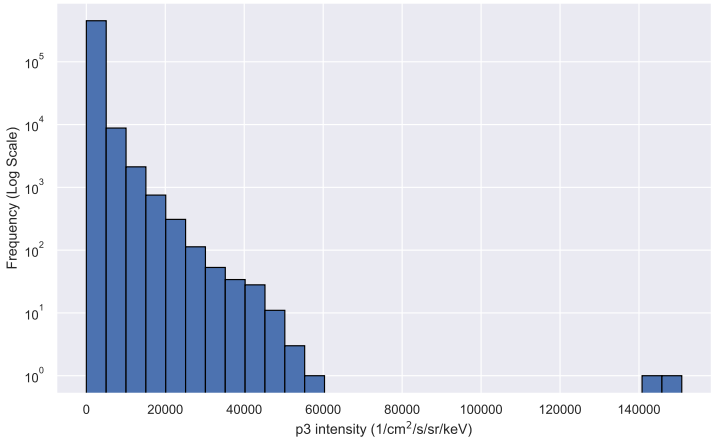


Figure A1. Histogram of the proton intensities measured by channel 3.

Appendix B Open Research

The authors express their gratitude to the team at the Cluster Science Archive (`https://csa.esac.esa.int`) for supplying the data. Additionally, we recognize the utilization of the OMNIWeb service and OMNI data from NASA/GSFC's Space Physics Data Facility (King & Papitashvili, 2005). The code and dataset used to derive the linear regression model can be found via the following link: `https://zenodo.org/records/10964236?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjkxMGQzY2EzLTMxNDAtNDZmNi05MjE5LWUzZmM1Y2I2OWM5MSIsImRhdGEiOnt9LCJyYW5kb20iOiIyMjQ1NjQzYTdlZjk3YzNjODA1MzdlZGJhMmQyMzg2MyJ9.oKgcCJTFE6KbvqqjXNh3wzfneL3XeY6meWb-XhbcKumOx0ztugSGFtvCaLblb3WAWOE5ccrkVEWZDnZ9vEs6EQ`.

Acknowledgments

The database used in this study was generated within the team led by Fabio Gastaldello on "Soft Protons in the Magnetosphere focused by X-ray Telescopes" at the International Space Science Institute in Bern, Switzerland. We are particularly grateful to Dr. Gastaldello for his invaluable feedback on our work, which greatly enhanced this research. EK and SM are supported by the German Research Foundation (DFG) under number KR 4375/2-1 within SPP "Dynamic Earth". EK is also supported by the DFG under number KR 4375/4-1. We are grateful to Dr. Andrew Read and Dr. Steven Sembay for their insightful suggestions which significantly enhanced the representation of spatial proton intensity distribution in our plots.

References

Banerjee, A., Bej, A., & Chatterjee, T. N. (2012). On the existence of a long range correlation in the geomagnetic disturbance storm time (dst) index. *Astrophysics and Space Science*, *337*, 23–32. doi: 10.1007/s10509-011-0836-1

Bellégo, C., Benatia, D., & Pape, L. (2021). Dealing with logs and zeros in regression models. *CREST - Serie des Documents de Travail*. doi: 10 .2139 / ssrn .3444996

Blandford, R. D., & Ostriker, J. P. (1978, April). Particle acceleration by astrophysical shocks. *Astrophys. J.*, *221*, L29-L32. doi: 10.1086/182658

Branduardi-Raymont, G., & Wang, C. (2022). The smile mission. In C. Bambi & A. Santangelo (Eds.), *Handbook of x-ray and gamma-ray astrophysics* (pp. 1–22). Singapore: Springer Nature Singapore. doi: 10.1007/978-981-16-4544-0_39-1

Branduardi-Raymont, G., Wang, C., Escoubet, C. P., Adamovic, M., Agnolon, D., Berthomier, M., ... Zhu, Z. (2018). *Smile definition study report* (ESA/SCI No. 1). European Space Agency. doi: 10.5270/esa.smile.definition_study_report-2018-12

Crooker, N. U., Eastman, T. E., & Stiles, G. S. (1979). Observations of plasma depletion in the magnetosheath at the dayside magnetopause. *Journal of Geophysical Research: Space Physics*, *84*(A3), 869-874. doi: 10.1029/JA084iA03p00869

Daly, P. W., & Kronberg, E. A. (2023). *User guide to the rapid measurements in the cluster science archive (csa)* (User Guide No. CAA-EST-UG-RAP 6.1). Max Planck Institute for Solar System Research. Retrieved 2023-11-05, from `https://www2.mps.mpg.de/dokumente/projekte/cluster/rapid/Rapid_Userguide.pdf`

Dorelli, J. C. (2019). Does the solar wind electric field control the reconnection rate at earth's subsolar magnetopause? *Journal of Geophysical Research: Space Physics*, *124*(4), 2668-2681. doi: 10.1029/2018JA025868

Fioretti, V., Bulgarelli, A., Malaguti, G., Spiga, D., & Tiengo, A. (2016). Monte carlo simulations of soft proton flares: testing the physics with xmm-newton. In J.-W. A. den Herder, T. Takahashi, & M. Bautz (Eds.), *Space telescopes and instrumentation 2016: Ultraviolet to gamma ray* (Vol. 9905, p. 99056W). SPIE. doi: 10.1117/12.2232537

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263. Retrieved 2023-10-07, from `http://www.jstor.org/stable/2841583` doi: 10.2307/2841583

Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., & Vasyliunas, V. M. (1994). What is a geomagnetic storm? *Journal of Geophysical Research: Space Physics*, *99*(A4), 5771-5792. doi: 10.1029/93JA02867

Hastie, T., Friedman, J., & Tibshirani, R. (2001). Linear methods for regression. In *The elements of statistical learning: Data mining, inference, and prediction* (pp. 41–78). New York, NY: Springer New York. doi: 10.1007/978-0-387-21606-5_3

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67. Retrieved from `https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634` doi: 10.1080/00401706.1970.10488634

Hubbard, M. W. J., Buggey, T. W., Hall, D., Feldman, C., Keelana, J., Hetherington, O., ... Holland, A. (2024). Techniques for estimating radiation damage from particles passage and focusing from micro pore optics. *Journal of Astronomical Telescopes, Instruments, and Systems*. (In press)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Linear regression. In *An introduction to statistical learning* (1st ed., p. 59-126). Springer New York, NY. doi: 10.1007/978-1-4614-7138-7

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, *110*(A2). doi: 10.1029/2004JA010649

Koga, D., Gonzalez, W. D., Souza, V. M., Cardoso, F. R., Wang, C., & Liu, Z. K. (2019). Dayside magnetopause reconnection: Its dependence on solar wind and

539 magnetosheath conditions. *Journal of Geophysical Research: Space Physics*,
540 *124*(11), 8778-8787. doi: 10.1029/2019JA026889
541 Kronberg, E. A., Clerc, N., Cros, A., de Plaa, J., Gastaldello, F., Gu, L., ... Valen-
542 tini, N. (2020). Prediction and Understanding of Soft-proton Contamination
543 in XMM-Newton: A Machine Learning Approach. *The Astrophysical Journal*,
544 *903*(2), 89. doi: 10.3847/1538-4357/abbb8f
545 Kronberg, E. A., Daly, P. W., Grigorenko, E. E., Smirnov, A. G., Klecker, B., &
546 Malykhin, A. Y. (2021). Energetic charged particles in the terrestrial magneto-
547 sphere: Cluster/rapid results. *Journal of Geophysical Research: Space Physics*,
548 *126*(9), e2021JA029273. doi: 10.1029/2021JA029273
549 Kronberg, E. A., Hannan, T., Huthmacher, J., Münzer, M., Peste, F., Zhou, Z.,
550 ... Ilie, R. (2021). Prediction of soft proton intensities in the near-earth
551 space using machine learning. *The Astrophysical Journal*, *921*(1), 76. doi:
552 10.3847/1538-4357/ac1b30
553 Kronberg, E. A., Kis, A., Klecker, B., Daly, P. W., & Lucek, E. A. (2009). Mul-
554 tipoint observations of ions in the 30–160 kev energy range upstream of the
555 earth's bow shock. *Journal of Geophysical Research: Space Physics*, *114*(A3).
556 doi: 10.1029/2008JA013754
557 Kronberg, E. A., Rashev, M. V., Daly, P. W., Shprits, Y. Y., Turner, D. L., Droz-
558 dov, A., ... Friedel, R. (2016). Contamination in electron observations of the
559 silicon detector on board cluster/rapid/ies instrument in earth's radiation belts
560 and ring current. *Space Weather*, *14*, 449-462. doi: 10.1002/2016SW001369
561 Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statisti-*
562 *cal models* (Fifth ed.). New York: McGraw-Hill Irwin.
563 Luhmann, J. G., Walker, R. J., Russell, C. T., Crooker, N. U., Spreiter, J. R., &
564 Stahara, S. S. (1984). Patterns of potential magnetic field merging sites on
565 the dayside magnetopause. *Journal of Geophysical Research: Space Physics*,
566 *89*(A3), 1739-1742. doi: 10.1029/JA089iA03p01739
567 Milan, S. E., Gosling, J. S., & Hubert, B. (2012). Relationship between interplan-
568 etary parameters and the magnetopause reconnection rate quantified from
569 observations of the expanding polar cap. *Journal of Geophysical Research:*
570 *Space Physics*, *117*(A3). doi: 10.1029/2011JA017082
571 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ...
572 Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of*
573 *Machine Learning Research*, *12*, 2825–2830.
574 Perinati, E., Freyberg, M., Yeung, M. C. H., Pommranz, C., Hess, B., Diebold, S.,
575 ... Santangelo, A. (2024). *Using srg/erosita to estimate soft proton fluxes at*
576 *the athena detectors.*
577 Raab, W., Branduardi-Raymont, G., Wang, C., Dai, L., Donovan, E., Enno, G.,
578 ... Zheng, J. (2016). Smile: a joint esa/cas mission to investigate the inter-
579 action between the solar wind and earth's magnetosphere. In J.-W. A. den
580 Herder, T. Takahashi, & M. Bautz (Eds.), *Space telescopes and instrumen-*
581 *tation 2016: Ultraviolet to gamma ray* (Vol. 9905, p. 990502). SPIE. doi:
582 10.1117/12.2231984
583 Raschka, S., Liu, Y., & Mirjalili, V. (2022). Predicting continuous target variables
584 with regression analysis. In *Machine learning with pytorch and scikit-learn :*
585 *develop machine learning and deep learning models with python* (p. 269-304).
586 Packt Publishing.
587 Read, A. M., & Ponman, T. J. (2003, October). The xmm-newton epic background:
588 Production of background maps and event files. *Astronomy & Astrophysics*,
589 *409*(1), 395–410. doi: 10.1051/0004-6361:20031099
590 Santosa, F., & Symes, W. W. (1986). Linear inversion of band-limited reflection seis-
591 mograms. *SIAM Journal on Scientific and Statistical Computing*, *7*(4), 1307-
592 1330. doi: 10.1137/0907087
593 Shue, J.-H., Chao, J. K., Fu, H. C., Russell, C. T., Song, P., Khurana, K. K., &

Singer, H. J. (1997). A new functional form to study the solar wind control of the magnetopause size and shape. *Journal of Geophysical Research: Space Physics*, *102*(A5), 9497-9511. doi: 10.1029/97JA00196

Sundberg, T., Haynes, C. T., Burgess, D., & Mazelle, C. X. (2016). Ion acceleration at the quasi-parallel bow shock: Decoding the signature of injection. *The Astrophysical Journal*, *820*(1), 21. doi: 10.3847/0004-637X/820/1/21

Tapping, K. F. (2013). The 10.7 cm solar radio flux (f10.7). *Space Weather*, *11*(7), 394-406. doi: 10.1002/swe.20064

Tsyganenko, N. A. (1995). Modeling the earth's magnetospheric magnetic field confined within a realistic magnetopause. *Journal of Geophysical Research: Space Physics*, *100*(A4), 5599-5612. doi: 10.1029/94JA03193

Walsh, B. M., Kuntz, K. D., Collier, M. R., Sibeck, D. G., Snowden, S. L., & Thomas, N. E. (2014). Energetic particle impact on x-ray imaging with xmm-newton. *Space Weather*, *12*(6), 387-394. doi: 10.1002/2014SW001046

Wang, C., Han, J. P., Li, H., Peng, Z., & Richardson, J. D. (2014). Solar wind-magnetosphere energy coupling function fitting: Results from a global mhd simulation. *Journal of Geophysical Research: Space Physics*, *119*(8), 6199-6212. doi: 10.1002/2014JA019834

Wilken, B., Axford, W. I., Daglis, I., Daly, P., Güttler, W., Ip, W. H., ... Ullaland, S. (1997). Rapid. In C. P. Escoubet, C. T. Russell, & R. Schmidt (Eds.), *The cluster and phoenix missions* (pp. 399–473). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-011-5666-0_14