# Prediction of Soft Proton Intensities in the Near-Earth Space Using Machine Learning

Elena A. Kronberg[1] , Tanveer Hannan[2,3] , Jens Huthmacher[2] , Marcus Münzer[2] , Florian Peste[2] , Ziyang Zhou[2] ,
Max Berrendorf[2] , Evgeniy Faerman[2] , Fabio Gastaldello[4] , Simona Ghizzardi[4] , Philippe Escoubet[5] , Stein Haaland[6,7,8] ,
Artem Smirnov[9,10] , Nithin Sivadas[11] , Robert C. Allen[12] , Andrea Tiengo[4,13,14] , and Raluca Ilie[15]

[1] Department of Earth and Environmental Sciences (Geophysics), Ludwig Maximilian University of Munich (LMU) Munich, Theresienstr. 41, Munich, D-80333, Germany; elena.kronberg@lmu.de
[2] Institute of Informatics, LMU Munich, Oettingenstraße 67, Munich, D-80538, Germany
[3] Department of Statistics, LMU Munich, Ludwigstraße 33, Munich, D-80539, Germany
[4] Instituto di Astrofisica Spaziale e Fisica Cosmica (INAF-IASF), Milano, via A. Corti 12, I-20133 Milano, Italy
[5] European Space Research and Technology Centre, Noordwijk, Keplerlaan 1, 2201 AZ, The Netherlands
[6] Birkeland Centre for Space Science, University of Bergen, Allégaten 55, NO-5007 Bergen, Norway
[7] Max Planck Institute for Solar System Research, Justus-von-Liebig-Weg 3, Göttingen, D-37077, Germany
[8] The University Centre in Svalbard, Longyearbyen, Svalbard, N-9171, Norway
[9] German Research Centre for Geosciences, Albert-Einstein-Straße 42–46, Potsdam, D-14473, Germany
[10] Institute of Physics and Astronomy, University of Potsdam, Karl-Liebknecht-Straße 24/25, Potsdam, D-14476, Germany
[11] Department of Electrical and Computer Engineering, Boston University, 8 Saint Mary's Street, Boston, MA 02134, USA
[12] Johns Hopkins University Applied Physics Lab, 11100 Johns Hopkins Road, Laurel, MD 20723, USA
[13] Scuola Universitaria Superiore IUSS Pavia, piazza della Vittoria 15, I-27100 Pavia, Italy
[14] INFN, Sezione di Pavia, via A. Bassi 6, I-27100 Pavia, Italy
[15] University of Illinois at Urbana-Champaign, 306 N. Wright Street, 5054 ECEB, Urbana, IL 61801, USA

## Abstract

The spatial distribution of energetic protons contributes to the understanding of magnetospheric dynamics. Based upon 17 yr of the Cluster/RAPID observations, we have derived machine-learning-based models to predict the proton intensities at energies from 28 to 962 keV in the 3D terrestrial magnetosphere at radial distances between 6 and 22 $R_E$. We used the satellite location and indices for solar, solar wind, and geomagnetic activity as predictors. The results demonstrate that the neural network (multi-layer perceptron regressor) outperforms baseline models based on the k-nearest neighbors and historical binning on average by ~80% and ~33%, respectively. The average correlation between the observed and predicted data is about 56%, which is reasonable in light of the complex dynamics of fast-moving energetic protons in the magnetosphere. In addition to a quantitative analysis of the prediction results, we also investigate parameter importance in our model. The most decisive parameters for predicting proton intensities are related to the location—Z geocentric solar ecliptic direction—and the radial distance. Among the activity indices, the solar wind dynamic pressure is the most important. The results have a direct practical application, for instance, for assessing the contamination particle background in the X-ray telescopes for X-ray astronomy orbiting above the radiation belts. To foster reproducible research and to enable the community to build upon our work we publish our complete code, the data, and the weights of trained models. Further description can be found in the GitHub project at https://github.com/Tanveer81/deep_horizon.

Unified Astronomy Thesaurus concepts: Astronomy data modeling (1859); Astronomy data analysis (1858); Space plasmas (1544); X-ray telescopes (1825); X-ray detectors (1815); X-ray observatories (1819)

## 1. Introduction

Understanding the distribution and dynamics of energetic protons in the near-Earth space is not just essential for magnetospheric physics. Energetic protons are also suspected to damage space-based instruments and to affect their scientific performance. For example, X-ray telescopes such as Chandra (Weisskopf et al. 2002) and the X-ray Multi-Mirror Mission (XMM-Newton, Jansen et al. 2001) are suffering from contamination by so-called soft protons (SP, De Luca & Molendi 2004; Kuntz & Snowden 2008; Leccardi & Molendi 2008). These are protons at energies in the range of tens of keV up to a few MeV. The SPs that populate the solar wind and the Earth's magnetosphere can damage CCD detectors, leading to a loss of available exposure time due to an increased background rate.

Consequently, the performance of future X-ray missions orbiting in the magnetosphere and the solar wind depends on how well the instruments are protected from the SP. For example, the Advanced Telescope for High-ENergy Astrophysics mission (Nandra et al. 2013) plans to deploy an array of magnets to deflect charged particles away from the instruments (Fioretti et al. 2018; Lotti et al. 2018). Moreover, the original orbit choice might be changed from L2 to L1 due to a better understanding of this region's energetic particle dynamics. The Solar wind Magnetosphere Ionosphere Link Explorer (SMILE, Raab et al. 2016) mission is also concerned with energetic particle levels in the magnetosheath that it will experience during its polar orbit.

There are few studies related to the energetic proton population in near-Earth space. In contrast to this work, they focus on well-confined regions around the Earth. For example, Meng et al. (1981), Kronberg et al. (2012), and

Kronberg et al. (2015) have studied the dependence of the distribution of energetic protons on solar wind and geomagnetic activity parameters in the plasma sheet. The plasma sheet is one of the few regions where it is expected to experience an enhanced background; this is also true for XMM; see, for example, Rosenqvist et al. (2002). In Kronberg et al. (2015) the distribution is projected to the equatorial plane. However, the missions mentioned above plan on taking observations at high latitudes.

Energetic protons can also be observed in other regions. The region upstream of the bow shock, for instance, is known to be populated with energetic protons due to effective acceleration at the bow shock (e.g., Lee 1982; Kronberg et al. 2009). Solar wind phenomena such as coronal mass ejections and corotational interaction regions are associated with energetic charged particles observed by spacecraft in the region upstream of the bow shock. The bow shock also serves to heat the solar wind plasma as it flows into the magnetosheath, the boundary region between the bow shock and the magnetopause, resulting in a harder energy spectrum. This region is generally considered void of energetic particles ($\sim$100 keV). However, occasionally, phenomena associated with energetic particles such as hot flow anomalies and plasma jets (e.g., Facskó et al. 2010; Savin et al. 2014) occur. Some of these energetic particles escape the magnetosphere (e.g., Kronberg et al. 2011). The diamagnetic cavities at cusps are also effective particle accelerators (e.g., Nykyri et al. 2012).

Studying each region in isolation enhances our knowledge about the physics in these regions and simplifies modeling. However, the boundaries between the regions are not always well defined, for instance, because of the quasi-parallel bow shock formation or Kelvin–Helmholtz instability. Therefore, it makes sense for space weather applications to study the proton intensities in the near-Earth region holistically. Moreover, most previous studies only consider a few input solar or geomagnetic parameters instead of utilizing the full range of them.

Kronberg et al. (2020) studied the dependence of SP contamination in the XMM-Newton telescope on location and various solar and geomagnetic parameters using a machine-learning approach. The study revealed the strongest dependence of the contamination on the location and the solar wind velocity. Simultaneously, parameters such as the AE or SYM-H indices (which measure the geomagnetic activity, namely, the disturbance of the magnetic field at high latitudes resulting from auroral electrojets (AE) in the northern hemisphere and at mid-latitudes, respectively) have shown significantly lower importance levels for contamination, which was rather unexpected from common knowledge of magnetospheric dynamics.

In this study we derive a predictive model for the energetic proton intensities using the Cluster mission observations in the near-Earth space environment. We exclude the region of the radiation belts since the proton intensity levels in this region are much higher than those in the outer magnetosphere. Our experience has shown that the model tends to predict only the intensities in the radiation belts if it is included. To enable modeling the complex non-linear multidimensional dependencies, we employ a machine-learning model instead of simple linear models.

To summarize, our study aims are to (1) test the capability of machine-learning algorithms to predict energetic particle populations in the near-Earth space, (2) reveal which parameters are the most important for the prediction of energetic protons at different energies, and (3) help future missions planning to deal best with the effects of SP.

## 2. Observations and Data Analysis

In this section, we describe which data we use and how it was obtained and preprocessed. The data can be found on Zenodo:doi:10.5281/zenodo.4718561.

### 2.1. Proton Observations

The Research with Adaptive Particle Imaging Detectors (RAPID) instrument (Wilken et al. 2001) on four Cluster satellites (Escoubet et al. 2001) measures distributions of energetic electron and ion intensities from $\sim$30 keV to $\sim$4 MeV. Around 50 data products are produced from the raw data and delivered to the Cluster Science Archive (CSA)[16] by the RAPID team. We chose to work with proton observations from spacecraft (SC) 4 (Tango), which has continuous observations from 2001 through the present day and can be combined with other ion measurements on Cluster. The omnidirectional energetic proton intensities can be found at CSA under the product `proton_Dif_flux__C4_CP_RAP_HSPCT` (Daly & Kronberg 2010). We took the first seven energy channels as the labels in our experiment, which represented the energy ranges $p1 = 28$–64 keV, $p2 = 75$–92 keV, $p3 = 92$–160 keV, $p4 = 160$–374 keV, $p5 = 374$–962 keV, $p6 = 962$–1885 keV, $p7 = 1885$–4007 keV, respectively. We exclude the region of the radiation belts (radial distance, `rdist < 6` $R_E$) from the data set.

For our experiments, we preprocessed the data as follows. First, we eliminated outliers, namely values below 0 and above $10^8$. Then, the data with original 4 s resolution were averaged over 1 minute because we have observed many rapid fluctuations within each minute and the predictors related to the solar and geomagnetic activity have the highest resolution of 1 minute. To be more precise, we calculated the mean proton intensities each minute for each energy channel and used the beginning of each minute as a time stamp.

Since the proton intensity values span multiple orders of magnitude, we use the common logarithms of the intensities as input to the model. However, since the proton intensity measurements can contain zero values, we cannot directly apply the logarithm. We investigated two methods: replacing the zero values with very small values, i.e., one-tenth of the smallest non-zero value, or dropping all measurements where the intensity is zero.

We obtained better results when dropping the zero values in the sense that, in this case, the model was more focused on predicting values above zero. In considering zero values, the model was skewed to the prediction of zero values as they are many. Since high proton intensities are dangerous for the performance of X-ray telescopes, we have decided to base our model on proton intensities above zero. Developing the model that focuses on zero/or close-to-zero proton intensities requires a separate study. Since the number of zero values differs across channels, we performed these operations independently for each channel. The highest energy channels $p6$ and $p7$ were dropped from the data set because they contain too many missing and zero values.

---

[16] https://csa.esac.esa.int

**Table 1**
Overview of Input Features Used and Their Units

| Feature | Unit | Description |
|---|---|---|
| x, y, z | $R_E$ | position of Cluster in GSE coordinate system |
| rdist | $R_E$ | radial distance from the Earth |
| BimfxGSE, BimfyGSE, BimfzGSE | nT | x, y, and z components of the interplanetary magnetic field (IMF) in GSE |
| VxSW_GSE, VySW_GSE, VzSW_GSE | km s$^{-1}$ | x, y, and z components of the solar wind speed in GSE |
| NpSW | n cm$^{-1}$ | solar wind density |
| Temp | K | solar wind temperature |
| Pdyn | nPa | solar wind dynamic pressure |
| AE_Index | nT | auroral electrojet index |
| SYM-H_index | nT | symmetric H-component index |
| F107 | sfu | the solar radio flux at 10.7 cm |

From 15:21:00 on 2001 January 9 to 09:57:00 on 2018 February 19 UT, 6,051,937 minutes of data in total matched these criteria. We list the predictors, also called features (the variables that are potentially capable of predicting the proton intensities), in Table 1 and discuss them below. Their distributions, along with the distribution of the proton intensities, are shown in Figure 9 in the Appendix.

## 2.2. Predictors

In this subsection we introduce the predictors that we divide into groups: related to location in space and related to the solar, solar wind, and geomagnetic activity.

### 2.2.1. Location in the Near-Earth Space

Each 1 minute averaged value of proton intensity is associated with a location in the geocentric solar ecliptic (GSE) coordinate system represented by parameters x, y, and z. The position coordinates were taken from the sc_r_xyz_gse__C4_C-P_AUX_POSGSE_1M auxiliary data set for SC4. We use them to obtain the radial distance from the Earth: the parameter rdist. Throughout the paper, distances are given in $R_E$ units. The distribution of the proton intensities and the number of their samples in the GSE system are shown in Figures 1 and 2. From Figure 2 one can see that the number of samples of proton intensity is strongly skewed toward the southern hemisphere and especially high latitudes. This is because the Cluster trajectories are inclined toward this hemisphere and cover well the high-latitude region. This is also the region of high interest for the XMM mission.

Figure 1 (left and middle panels) shows high intensities around the equatorial plane, ZGSE = 0, namely the plasma sheet at the night side and the region on the closed magnetic field lines at the dayside in the XZ and YZ planes. There are also mid-altitude cusps with less intensive plasma around XGSE = 0, in the same planes, above and below $\sim$3 $R_E$ and $\sim$−3 $R_E$ in ZGSE direction, respectively. Enhanced proton intensities at XGSE $\simeq$ 3–6 $R_E$ at higher latitudes than the plasma sheet (above and below $\sim$5 $R_E$ and $\sim$−5 $R_E$ in ZGSE direction, respectively) in the XZ plane are also visible. This intensity increase, at least partially, can be explained by the presence of the diamagnetic cavities near the cusps which can trap the protons and lead to their effective acceleration (see, e.g., Nykyri et al. 2011, 2012). The intensities are higher in the northern hemisphere than in the southern

hemisphere. The dusk–dawn distributions (in the YZ plane) show asymmetry with higher intensities at the northern hemisphere's dusk side. The same higher intensity at dusk is visible in the XY plane. This asymmetry agrees with observations of energetic protons >274 keV by Kronberg et al. (2015) and Luo et al. (2017). The proton intensities are higher on the dayside in the region on closed magnetic field lines.

In Figure 3, we plot the mean proton intensities versus individual predictors. In the following, for brevity, we refer to the logarithm of proton intensities simply as proton intensities. In panel (a), we observe a strong, almost linear decrease of the proton intensities with radial distance. The intensities almost linearly decrease with the ZGSE coordinates in the southern hemisphere (see panel (b)), while the dependence looks more complicated in the northern hemisphere. However, if we consider the distances between $\sim$−15 and $\sim$8 $R_E$ the maximum of proton intensity is shifted toward positive ZGSE = $\sim$2 $R_E$. Taking this value as the symmetry axis, the proton intensity distribution in the northern hemisphere also falls roughly linearly with the distance. Panel (c) shows that the maximum of the proton distribution is around $\pm$ 5 $R_E$. The proton intensity falls rapidly at $\sim$−15 $R_E$ at the dawn side and $\sim$13 $R_E$ at the dusk side. The dependence of proton intensities on XGSE is rather complicated; see panel (d). The strong increase of the proton intensities at XGSE < −15 $R_E$ can be explained by reduced spacecraft sampling in the northern lobe at these distances; see the XZGSE plane in Figure 2. It falls strongly at $\sim$10 $R_E$ as expected at the magnetopause boundary, as also seen in Figure 1. These spatial dependencies resemble those observed for the SP contamination at the XMM-Newton X-Ray telescope in Kronberg et al. (2020); see also Ghizzardi et al. (2017).

### 2.2.2. Solar, Solar Wind, and Geomagnetic Activity

The proton intensities are joined with simultaneous observations of solar, solar wind, and geomagnetic parameters from the OMNI database[17] (King & Papitashvili 2005). The solar wind measurements are propagated to the bow shock of the terrestrial magnetosphere. We used the following parameters to characterize the SW: the proton density, NpSW [cm$^{-3}$]; components of the velocity (VSW) in the GSE coordinates, VxSW_GSE, VySW_GSE and VzSW_GSE [km s$^{-1}$]; the temperature, Temp [K] and components of the IMF in the GSE coordinates, BimfxGSE, BimfyGSE, and BimfzGSE [nT]. We also included the dynamic pressure, Pdyn [nPa], calculated as NpSW*VSW$^2$ × 1.67 · 10$^6$.

The solar wind velocity $V_y$ and $V_z$ components show that the deviation from the radial direction leads to an increase of the proton intensities, excluding cases with substantial deviation (>100 km s$^{-1}$) in positive Y- and ZGSE directions for which a decrease is observed; see Figure 3 (e). The proton intensities increase with the solar wind speed in the anti-sunward direction, $V_x$, and the temperature; see panels (f) and (g). The dependencies of the proton intensities on solar wind density (panel (h)) and the solar wind dynamic pressure (panel (i)) have non-linear trends. The change of the IMF's direction toward stronger absolute values mainly leads to increased proton intensities (see panel (j)).

Solar irradiance, which correlates well with the number of sunspots and characterizes the solar activity, influences the amount of protons in the near-Earth space. The solar irradiance is
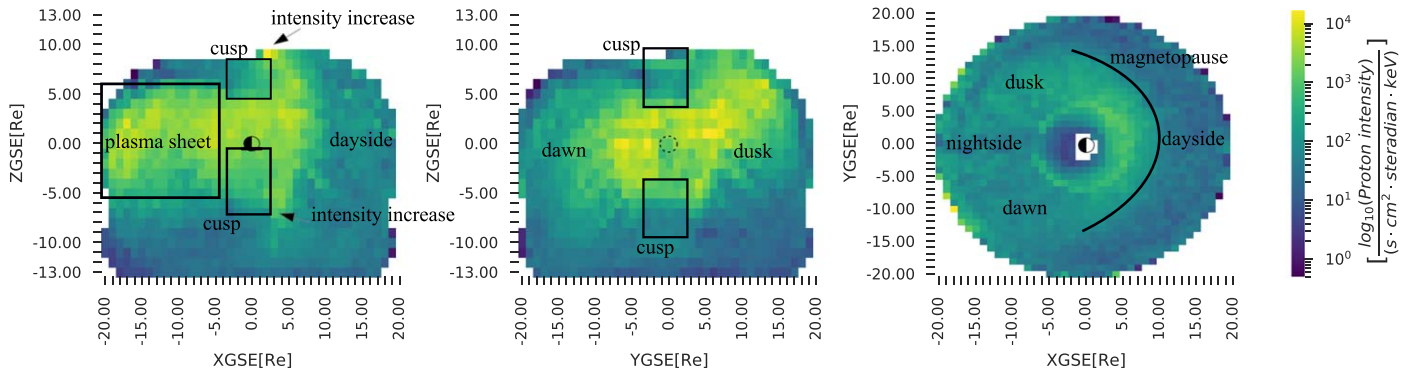
---

[17] https://omniweb.sci.gsfc.nasa.gov

**Figure 1.** Distribution of the observed proton intensity for the energy channel *p1* by SC4 from January 2001 to February 2018 in the GSE coordinate system. Resolution (bin size) is 1 $R_E$. Half black and half white circles, as well as the dashed circle, indicate the location of the Earth and are not to scale. Here zero values of proton intensity are replaced with small values.
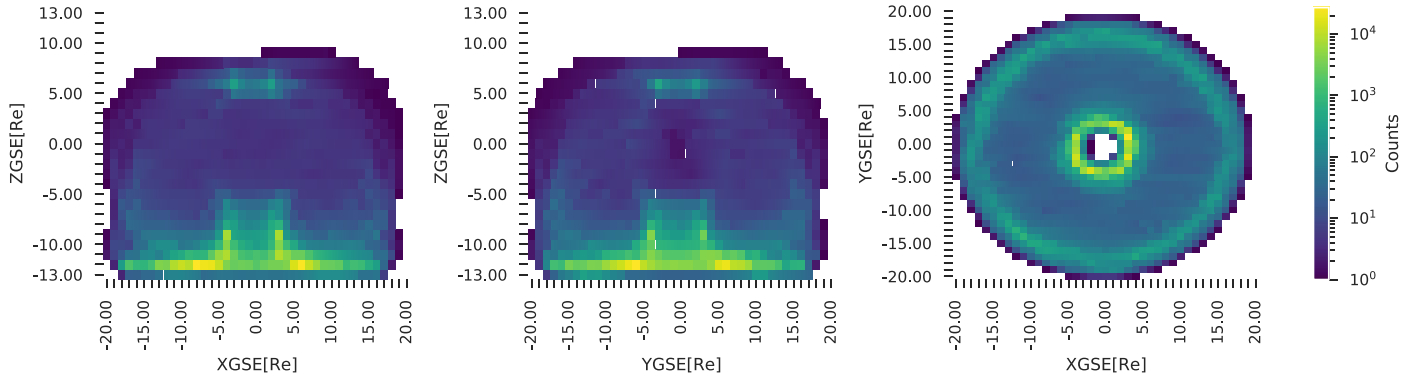


**Figure 2.** Distribution of the number of proton intensity measurements for the energy channel *p1* in the GSE coordinate system in the same format as Figure 1. Resolution (bin size) is 1 $R_E$.

represented by the radio flux measurement at 10.7 cm (2.8 GHz) called the F10.7 index, here `F107` [sfu] (Tapping 2013). Unlike many other solar indices, this parameter is measured reliably under any weather conditions at the Earth. The solar irradiance is non-linearly related to the proton intensities; see Figure 3 (k).

A parameter of geomagnetic activity such as the AE index, denoted as `AE_index`, in nT, characterizes the magnetic field disturbance in the auroral region of the northern hemisphere (Nosé et al. 2017). The relation of the proton intensities with the AE index is also non-linear; see Figure 3 (l). However, a general trend of increase in the proton intensities with the AE index is visible, namely with the geomagnetic activity at high latitudes. Another parameter related to the geomagnetic activity is the SYM-H index, denoted as `SYM-H` and measured in nT (Iyemori et al. 1992). This parameter characterizes the disturbance of the geomagnetic field at the equatorial regions. The geomagnetic activity related to the geomagnetic storms, characterized by the SYM-H index, shows a non-linear relation with proton intensities; see Figure 3 (m).

If we compare the trends of the proton intensity changes with the solar, solar wind, and geomagnetic parameters with those for the SP contamination from Kronberg et al. (2020) we notice general agreement between those. More details on the comparison are in Section 5.

### 2.3. Cross-correlations between Proton Intensities and Predictors

Figure 4 shows the Pearson correlation between parameters possibly related to proton dynamics. The correlation values range between $-1$ and 1. Values close to $-1/1$ mean perfect linear anticorrelation/correlation and values close to 0 mean the absence of linear correlation. One has to be careful in the interpretation of the Pearson correlation coefficient because it assesses only linear relationships. The proton intensities from channel *p1* are well correlated with the radial distance and the ZGSE location of observation, in agreement with Figure 3. From the OMNI parameters, the proton intensities of channel 1 are best linearly correlated with the `VxSW_GSE`, the same as the SP contamination in Kronberg et al. (2020).

### 2.4. Data Split

The full data set, as was mentioned above, comprises in total 6,051,937 measurements from 2001 January 09 15:21:00 UT to 2018 February 19 09:57:00 UT. We split the data set into a training (or development) set (80%) and a test set (20%). To prevent test leakage, we do not shuffle the data but split it by a time point with the original order preserved. Afterward, we additionally split the development data into a train (80%) and validation set (20%) again by time. We utilize the validation set to optimize the model hyperparameters.

We normalized the features by subtracting the median and dividing by the inter-quartile range.[18]

Table 2 summarizes the sizes and periods of the data subsets after performing preprocessing and splitting.

---

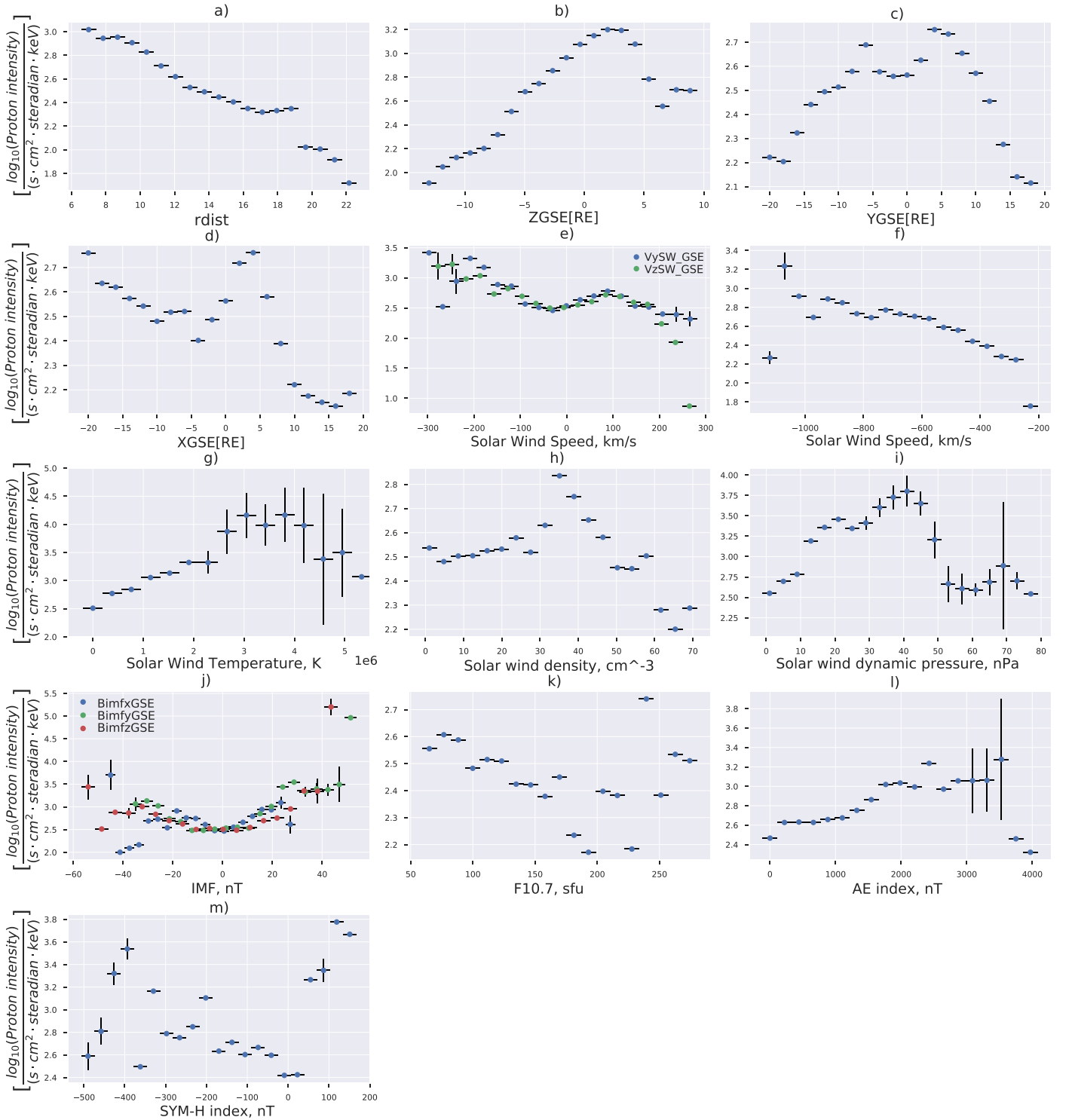[18] As implemented by `RobustScaler` in `sklearn`.

**Figure 3.** Relations of mean proton intensities for the energy channel 1 and (a)–(d) radial distance, `rdist`, ZGSE, YGSE, and XGSE, respectively; (e) the solar wind $V_y$ and $V_z$ components; (f) the solar wind radial velocity, $V_x$; (g)–(i) solar wind temperature, density, and dynamic pressure, respectively; (j) IMF components in GSE; (k) F10.7 parameter; (l) AE index and (m) SYM-H index. Vertical lines represent confidence intervals at 95% confidence level. Here zero values of proton intensity are replaced with small values.

## 3. Machine-learning Model for Proton Intensities

The relation between the proton intensities and the individual predictors discussed above is mostly non-linear; see Figure 3. Therefore, often a group of predictors or their ensemble help to obtain better predictions than the best individual predictor (Geron 2019).

We interpret the prediction of proton intensities as a regression problem of form

$$y \approx f([\boldsymbol{p}; \boldsymbol{x}_t]|\boldsymbol{\theta})$$

where $y$ is the proton intensity for a single channel, $\boldsymbol{p}$ is the spatial position, and $\boldsymbol{x}_t$ the additional input parameters as given
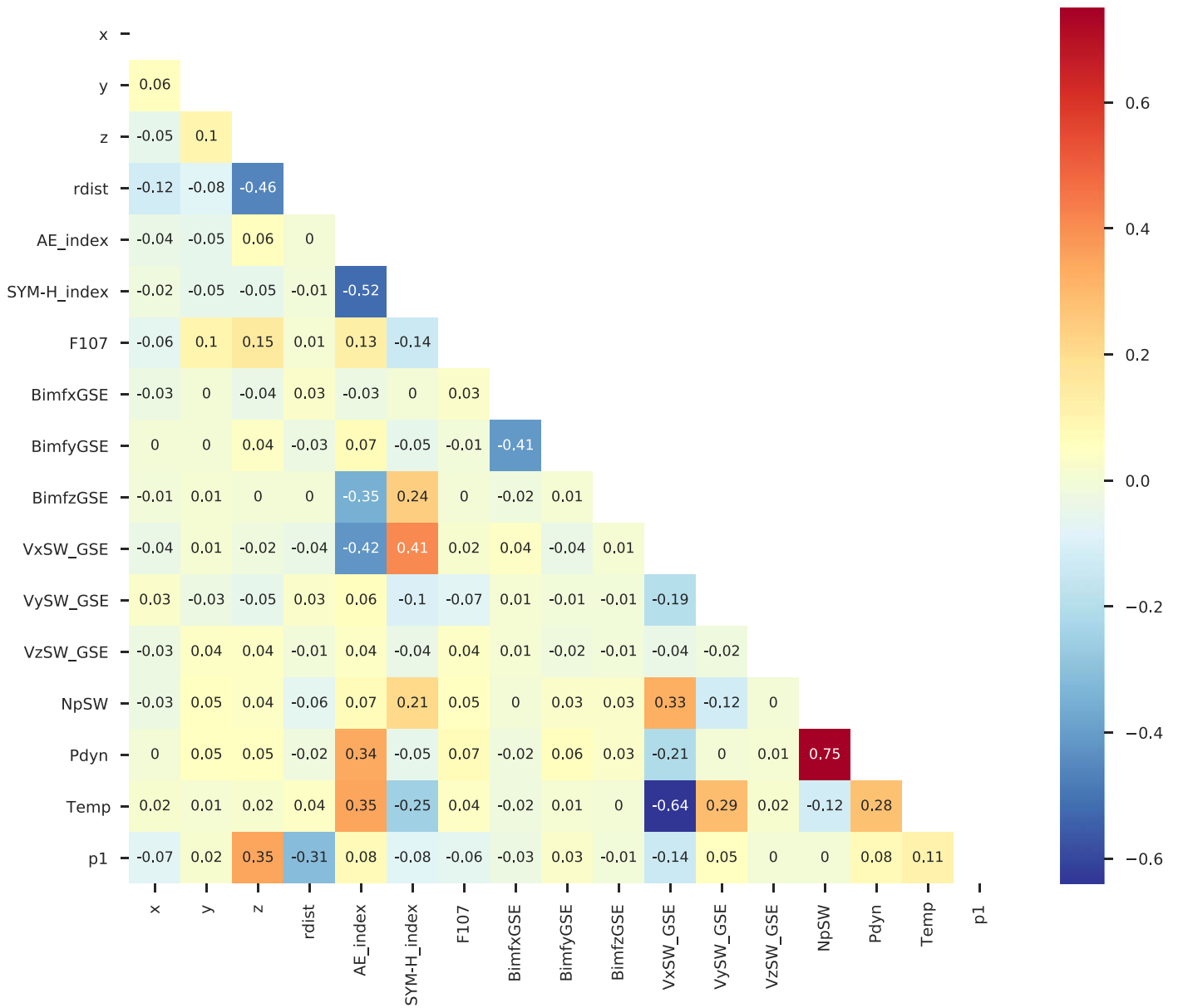
**Figure 4.** Pearson correlation matrix between input parameters and proton intensity (here, zero values are replaced with small values). For readability, the values are rounded to the second decimal.

**Table 2**
Size and Periods of the Data Subsets after Splitting.

| Subset | Start (UT) | End (UT) | Number of Data Points without Zero Values | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| | | | p1 | p2 | p3 | p4 | p5 |
| Train | 2001-01-09 15:21:00 | 2011-08-23 08:22:00 | 2,130,927 | 978,828 | 1,396,154 | 1,093,985 | 522,122 |
| Validation | 2011-08-23 08:24:00 | 2014-07-24 22:44:00 | 532,731 | 244,707 | 349,038 | 273,496 | 130,530 |
| Test | 2014-07-24 22:45:00 | 2018-02-19 09:57:00 | 694,867 | 320,363 | 447,790 | 329,581 | 131,208 |

in Table 1 at time $t$. [ ; ] denotes the concatenation and $\theta$ are model parameters that are to be estimated from the given data set. We study solving this regression task by applying various machine-learning models from which we select according to validation performance. The models' hyperparameters are also tuned according to validation performance.

We investigated the following linear models: linear regression (Galton 1886) with and without $l_1/l_2$ regularization

(lasso (Santosa & Symes 1986)/ridge regression (Hoerl & Kennard 1970)), least angle regression (Tibshirani et al. 2004) with and without $l_1$ regularization (lasso LARS regression), and linear support vector regression (SVR) (Cortes & Vapnik 1995).

We considered the following tree-based ensemble models: regression trees (Breiman et al. 1984), random forests, AdaBoost (Freund & Schapire 1995), ensembles of extremely randomized trees (extra trees; Geurts et al. 2006), gradient

**Table 3**
Validation Result of all Machine-learning Models Represented by a Spearman Correlation between Results Predicted by Different Models and Observed Proton Intensity for All Considered Energy Channels

| Model | p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|---|
| MLP | 61.4% | 50.4% | 46.1% | 37.2% | 18.5% |
| LGBM | 57.1% | 46.3% | 47.0% | 36.9% | 17.6% |
| AdaBoost | 56.9% | 53.9% | 52.5% | 34.7% | 23.2% |
| ExtraTrees | 53.3% | 51.3% | 50.9% | 41.6% | 9.7% |
| LinearSVR | 51.9% | 52.9% | 51.2% | 44.5% | 22.2% |
| RidgeRegression | 49.9% | 54.6% | 51.2% | 45.6% | 25.1% |
| LARSRegression | 49.9% | 54.6% | 51.2% | 45.6% | 25.2% |
| RandomForest | 39.4% | 34.4% | 34.0% | 30.4% | −9.6% |

**Table 4**
Test Result of All Machine-learning Models Represented by a Spearman Correlation between Results Predicted by Different Models and Observed Proton Intensity for All Considered Energy Channels

| Model | p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|---|
| MLP | 56.9% | 58.8% | 53.5% | 56.2% | 53.5% |
| LGBM | 56.7% | 60.2% | 60.3% | 60.3% | 57.7% |
| AdaBoost | 57.0% | 60.5% | 58.3% | 58.4% | 55.8% |
| ExtraTrees | 42.2% | 47.2% | 42.3% | 43.9% | 43.3% |
| LinearSVR | 46.5% | 52.6% | 50.5% | 51.3% | 44.6% |
| RidgeRegression | 44.3% | 52.8% | 51.2% | 51.6% | 40.3% |
| LARSRegression | 44.2% | 52.8% | 51.2% | 51.6% | 40.3% |
| RandomForest | 35.1% | 43.4% | 31.6% | 40.7% | 9.8% |

boosting (Friedman 2001), and light gradient boosting machines (LGBM; Ke et al. 2017).

We also used a multi-layer perceptron (MLP) regressor (Rosenblatt 1958). This regressor is a neural network with an input layer, one or more hidden layers, and one output layer. Each layer is a dense layer, $x^{(i+1)} = \sigma(Wx^{(i)} + b)$, where $W \in \mathbb{R}^{d_i \times d_{i+1}}$, $b \in \mathbb{R}^{d_{i+1}}$ are trainable parameters, and $\sigma$ is a non-linear activation function. The weights are trained *end-to-end* with an iterative (stochastic) gradient descent method using the gradient of the loss function's output with respect to parameters. The gradients can be efficiently computed using back-propagation. Important hyperparameters are the choice of the network structure, i.e., the number of layers and the hidden dimensions $d_i$.

## 4. Machine Learning Experiments

### 4.1. Setup

We used LGBM implementation from the LightGBM library (Microsoft Corporation 2008) and sklearn (Buitinck et al. 2013) for all other models. In order to evaluate the models' performance, we use the following measures: Spearman correlation, Pearson correlation, mean absolute error, mean squared error, and coefficient of determination ($R^2$). For model selection and hyper-parameter optimization, we focus on the Spearman correlation.

To demonstrate the necessity of using advanced machine-learning models, we compare the performance against two simple baselines: historical binning and k-nearest neighbors (kNN; Altman 1992).

In the historical binning baseline, we create spatial bins of training data with the k-means algorithm (Lloyd 1982) applied to the position features only. The number of bins is chosen independently for each channel based on validation performance.

**Table 5**
Spearman Correlation between Results Predicted by Different Models and Observed Proton Intensity for All Considered Energy Channels on Both Train and Test Split.

| Model | Split | p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|---|---|
| kNN | Test | 29.1% | 32.3% | 32.3% | 32.7% | 30.0% |
| HistBin | Test | 39.3% | 44.0% | 40.3% | 42.3% | 44.0% |
| MLP | Train | 73.5% | 72.6% | 70.6% | 70.8% | 74.8% |
| MLP | Test | 56.9% | 58.8% | 53.5% | 56.2% | 53.5% |

**Note.** The MLP model was chosen as the best configuration according to the validation Spearman correlation of channel *p1*.

For a test point, we determine the corresponding bin and use the average proton intensity of that corresponding region from the train data as the prediction.

In the kNN baseline, for a test data point, we determine the k-nearest neighbor that exists in the train set based on the Minkowski distance and interpolate its proton intensity for the position. Then we take the interpolated intensity at the test position as the prediction.

### 4.2. Results

In order to determine the best hyperparameters, we used random search over a pre-defined search space (see Table 6 in the Appendix) due to its higher sample efficiency than grid search (Bergstra & Bengio 2012). In this search, we trained numerous models on the training data. We utilized the ASHA Scheduler (Li et al. 2020) for its parallelism and extensive early stopping capabilities via the `ray tune` (Liaw et al. 2018) framework. After training, we evaluate each model on the validation data, and we choose the configuration that obtains the best validation Spearman correlation. To reduce resource consumption, we perform this optimization only for channel *p1* and apply the same best hyperparameters for the other channels, cf. Table 6 in the Appendix.[19] We obtain the best validation performance for an MLP model based on the value of *p1* channel for which the optimization was done and we select this choice; see Table 3. The second and the third best results based on these metrics are obtained for LGBM and AdaBoost models. If we calculate the average performance for all energy channels, the MLP model was outperformed by LinearSVR, LARSRegression, RidgeRegression, and Ada-Boost (listed in decreasing performace order). The test results are also shown for all machine-learning models; see Table 4. They indicate that MLP is outperformed by the LGBM and AdaBoost models considering the average performance of all energy channels. These results show that top performing machine-learning algorithms lead to relatively similar prediction rates and can be interchangeably used. Table 5 shows the final results for the MLP model on the training and test sets. The MLP model outperforms the baseline models such as kNN and historical binning by about 80% and 33%, respectively. Because the MLP model performs better on test data than all baselines, we conclude that it generalizes to unseen data. Note that a gap between model performance on training and test data is often observed for complex models and it is not necessarily an indication of overfitting. The average Spearman correlation between the observed and predicted data is about 56%. This

---

[19] We publish the trained models on Zenodo: doi:10.5281/zenodo.4593065.

value is reasonable considering the stochasticity and complex dynamics of the fast-moving energetic protons in the terrestrial magnetosphere.

Figure 5 shows the distributions of the observed proton intensities versus the predicted values for five energy channels. The left-hand panels show the training set distributions, whereas the right-hand panels represent the test set. Observed and predicted data for the training and test data sets agree relatively well. The data are mainly concentrated along the black dashed line, corresponding to a perfect correlation. The prediction of proton intensities in all five energy channels has a similar quality.

In Figure 6, we show a qualitative example of the model's predictions. The chosen time frame shows a plasma sheet crossing, the approach of the radiation belts, and the cusp region at the night side of the magnetosphere. The model predicts the proton intensities for the different energy channels, rarely deviating more than one order of magnitude, which is considered a good prediction for energetic particle dynamics. From the beginning of this time interval until ∼13 UT, substorm processes (AE > ∼ 150 nT) were observed. The SYM-H index was above −30 nT, excluding the time period from ∼08 to 09 UT during which it dropped almost down to −40 nT, indicating that during this interval there was almost no magnetic storm activity. Figure 7 shows another example that demonstrates the model performance during a magnetic storm with the SYM-H index dropping down to −90 nT. During this time interval, the Cluster spacecraft was in the near-Earth plasma sheet. As for the previous event, the model predicts the proton intensities mainly within an order of magnitude. We must, however, mention that our model is not developed specifically for the prediction of the proton intensities during magnetic storms which are rather rare events. Therefore, prediction efficiency during geomagnetic storms is not expected to be high.

To evaluate the statistical significance of our results, we use a two-sided hypothesis test on the Spearman correlation. The null hypothesis states that the predictions are uncorrelated to the observations. For all five channels, we obtain a $p$ value of 0, or, more correctly, below the smallest positive number in float64. Hence, we can reject the null hypothesis, i.e., the predictions are correlated to the observations. Thus, we can deduce from this result that our model learned the trend of proton intensity.

### 4.3. Feature Importances

To investigate the importance of the individual input features on the model's prediction, we utilize permutation feature importance (Fisher et al. 2019) since it is model-agnostic and interpretable. For an investigated feature column, e.g., $x$, its values are shuffled to break any association with the target value, i.e., the proton intensity. Then, we re-calculate the model's prediction and evaluate the Spearman correlation between the prediction and the original target values. The feature importance is measured as the decrease in performance: thus, a more substantial decrease in model performance corresponds to a more important feature. Negative feature importance is also possible when the prediction improves when breaking an input feature's association. We use bootstrapping with 100 iterations and a subsampling fraction of 75% to estimate the mean and variance of this statistic to assess significance.

Figure 8 shows the feature importance for each input variable and channel. The parameters related to the location show significantly higher importance than parameters related to solar, solar wind, and geomagnetic activities. From those, on average, the most vital dependence is seen for ZGSE and the radial distance. In the ZGSE direction, the satellite crosses more different magnetospheric regions, e.g., plasma sheet, cusp, lobe, and magnetosheath, than in other directions. This leads to stronger gradients of the proton intensities in this direction; see also Figure 1. This, at least partially, explains the strongest dependence on this parameter. The least important location parameter is YGSE. From the other parameters, the solar wind dynamic pressure is the most important parameter for predicting the proton intensities, followed by, on average, the AE index.

## 5. Discussion

We want to note that the most substantial linear dependence of the proton intensities among the OMNI parameters is on the VxSW_GSE, cf. the Pearson correlation in Figure 4. This linear dependence is also evident in Figure 3. However, the feature importance derived in our model indicates Pdyn and the AE index as the most important OMNI parameters; cf. Figure 8. In Figures 3 and 4, we always consider one input variable and one output variable in isolation. The model, however, is not restricted to do so, but rather can combine features to get "latent features", i.e., combinations of individual features or their interaction.

The proton intensities' dependency on different parameters is generally very similar to the dependencies of the XMM SP contamination count rates derived in Kronberg et al. (2020) despite differences in the trajectories of the spacecraft. For example, a clear dependence on rdist is similar in enhancing the proton intensities and count rates between 17 and 19 $R_E$. For the YGSE dependency, we observe for both proton intensities and SP count rates a decrease at YGSE $\simeq 0\,R_E$ and dusk-ward asymmetry at distances YGSE $\simeq \pm 10\,R_E$. For the XGSE dependency, we observe high proton intensities and SP count rates between XGSE $\sim 0$ and 5 $R_E$. Similar dependencies are observed for the almost linear relation with VxSW_GSE up to $\sim -900$ km s$^{-1}$, for the Temp, Np up to 20 cm$^{-3}$, BimfxGSE and BimfyGSE up to |10| nT, AE index up to $\sim 700$ nT and the SYM-H index within the range between −150 and 50 nT. For Pdyn, the considered scale ranges are very different, and, therefore, the resolution is different. However, we still see an average increase of Pdyn with the proton intensities/SP count rates at least up to 6 nPa. The only considerable disagreement for the relations of the proton intensities and parameters is observed for the F10.7 index. For example, for the XMM count rates, one observes an increase with the F10.7 index up to 150 sfu. However, the proton intensities observed by Cluster show a decrease. We can draw a relation to the complexity of the processes associated with the solar cycle or/and to the bias related to the spacecraft trajectory during the solar cycle. The Cluster mission was located closer to the Earth during the years related to the 24th solar cycle (higher proton intensities versus lower F10.7 index) which is significantly weaker than the 23rd solar cycle during which the XMM measurements were done (lower proton intensities versus higher F10.7 index). The highest Pearson correlation derived for the proton intensities/SP count rates are the ZGSE, rdist, and VxSW_GSE in both cases. In conclusion, based on
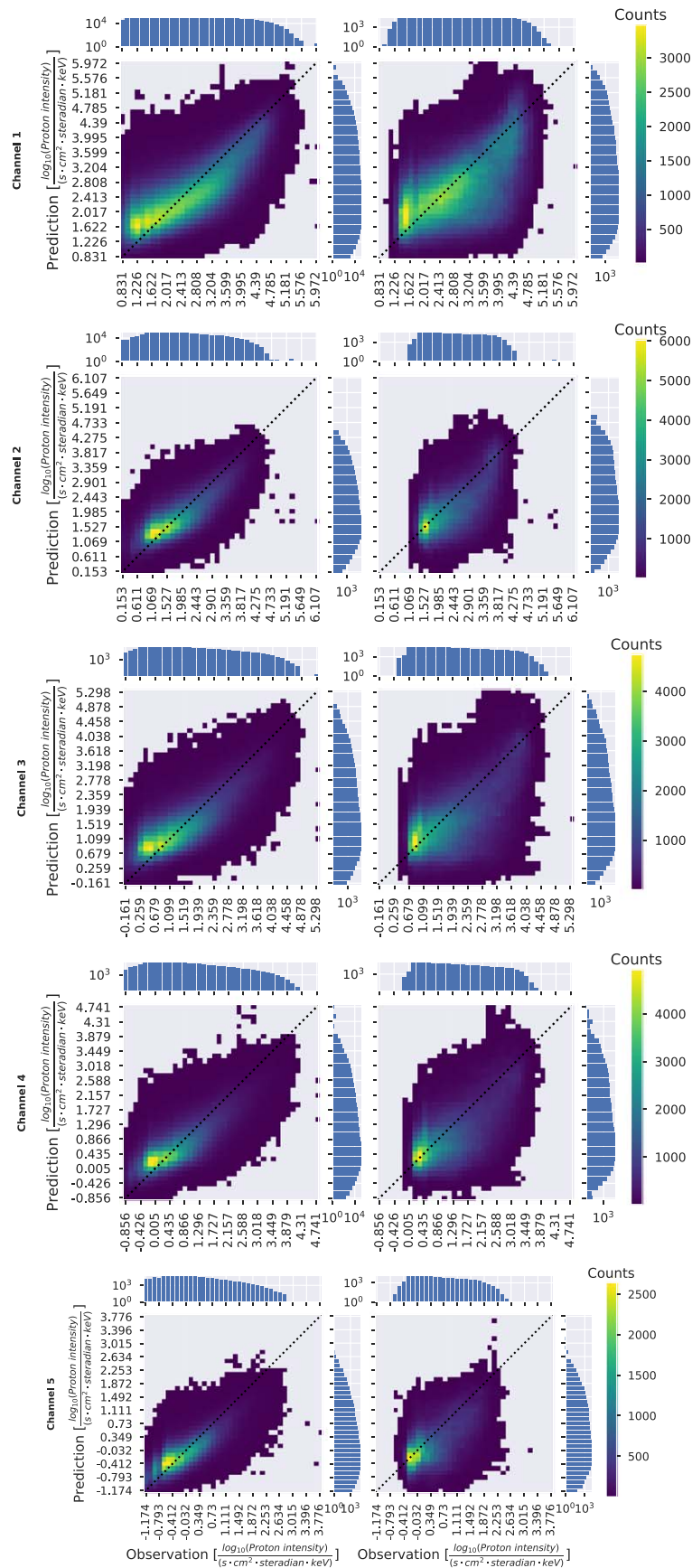
**Figure 5.** The observed (*x*-axis, in this case zero values are dropped) vs. the predicted (*y*-axis) proton intensities for the energy channels *p1* to *p5* from the training (left) and the test (right) set. The color represents the number of samples in the corresponding bin, where a good model will predict most of the intensities along the diagonal, i.e., closely matching the measurements. The underlying model for those plots is the best MLP determined according to the validation Spearman correlation for channel *p1*.
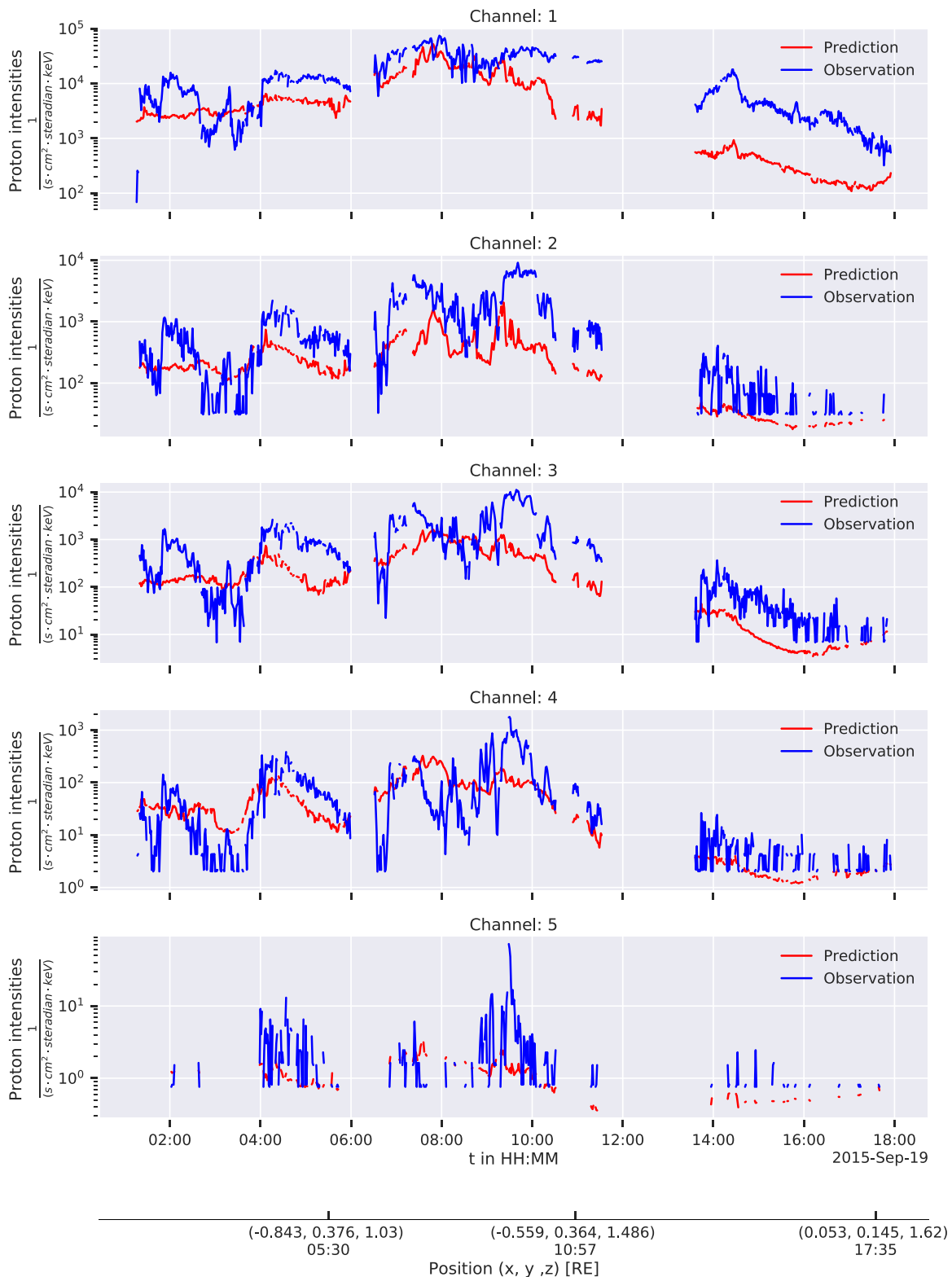
**Figure 6.** Predicted intensities (red) vs. measured intensities (blue, in this case; zero values are dropped) within the time interval on 2015 September 19 from 01:15 to 20:30 UT.

Figures 3 and 4 our energetic proton observations indicate that it is likely that they could be a cause of the XMM SP contamination. The physical processes leading to proton energization associated with high solar wind speed are related, for example, to the formation of the quasi-parallel shock, acceleration processes associated with reconnection at the day

side, magnetospheric storms, and substorms; see more detailed discussion in Kronberg et al. (2020).

The dependencies of proton intensities and SP count rates on the different parameters are generally similar. A notable difference is that `Pdyn` and the AE index had the highest importance of the proton intensities in the MLP model. In
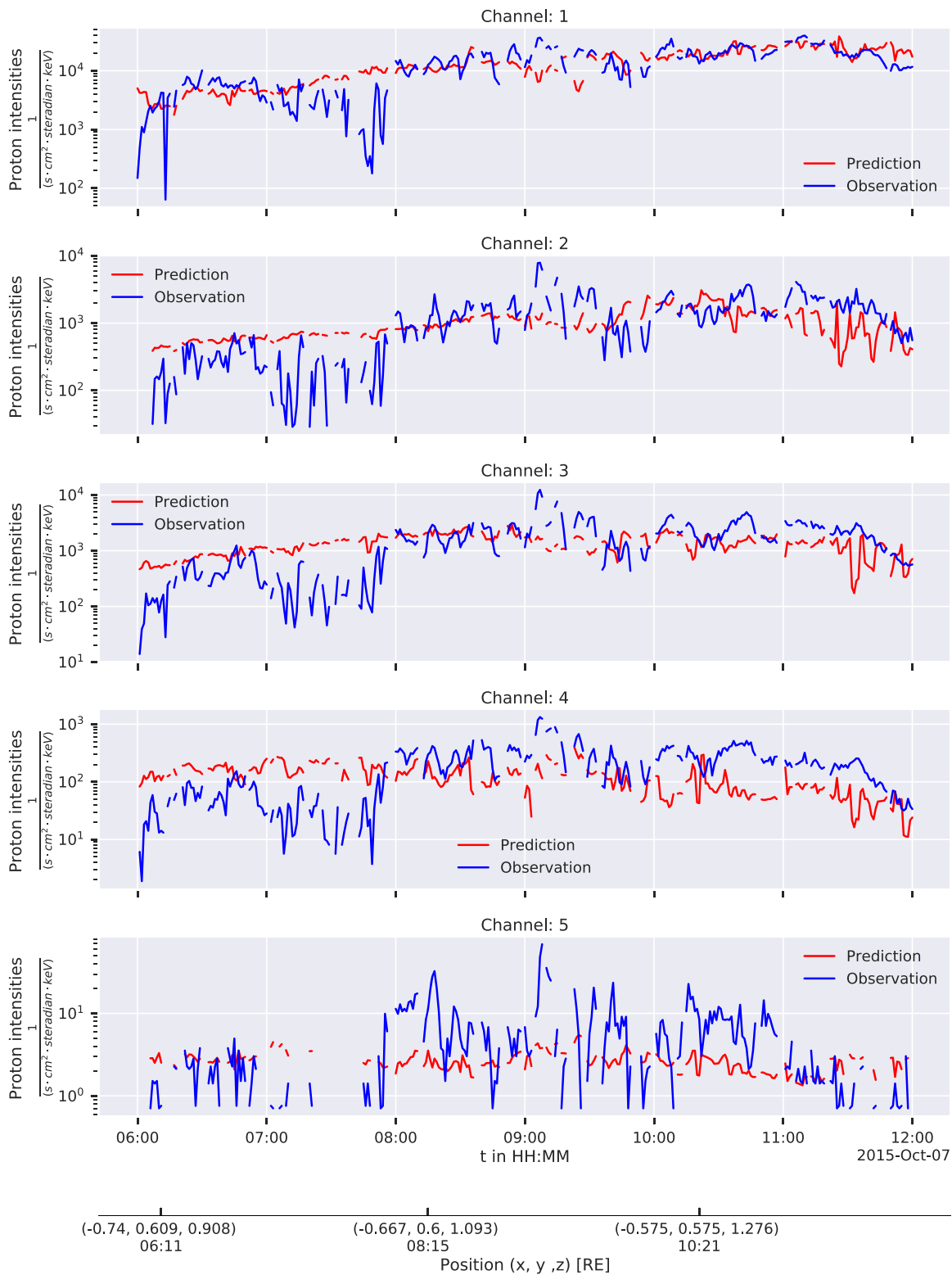
**Figure 7.** Predicted intensities (red) vs. measured intensities (blue, in this case; zero values are dropped) within the time interval on 2015 October 07 from 06:00 to 12:00 UT. During this time interval, a magnetic storm was observed with SYM-H index dropping almost down to −90 nT.

contrast, the SP count rates in the extra random forest model had the highest importance for the `VxSW_GSE` and F10.7 index (from the solar, the solar wind, and geomagnetic parameters). This difference most likely occurs because of using different non-linear models and different observation lengths (one versus one and a half solar cycles). Another possible reason is that the

Cluster trajectories cover the plasma sheet much more thoroughly than those of the XMM. The plasma sheet dynamics is strongly related to substorm activity indicated by the AE index and is strongly dependent on `Pdyn`. In the study by Smirnov et al. (2019) based on Cluster observations, the dependence of the electron intensities at L-shells between 4 and
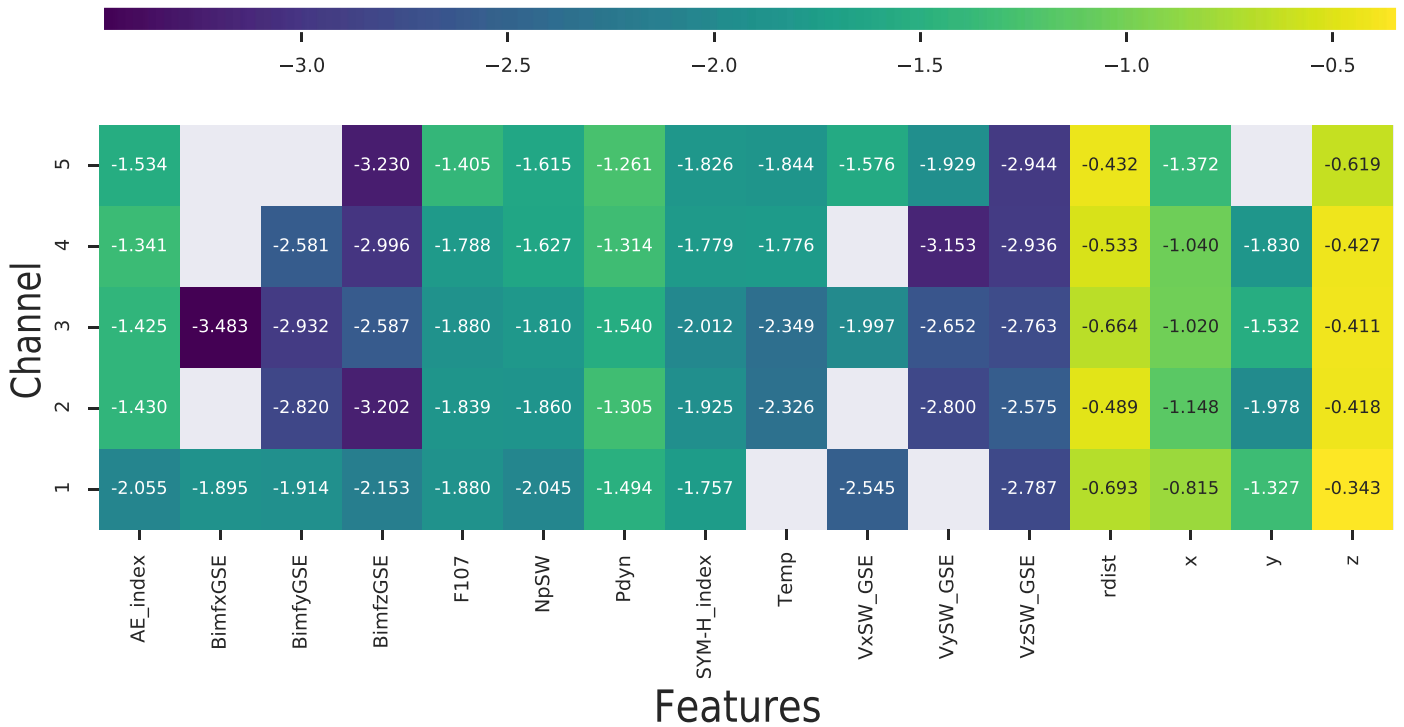
**Figure 8.** Permutation feature importance in the logarithmic (base 10) scale calculated for the MLP model on the test set. Original negative values of the feature importance are shown in gray color.

6 for ∼1.5 solar cycle (approximately the same period as in our study) was also the best correlated with the Pdyn and AE index. The electron intensities in the outer radiation belts are governed by the substorm associated activity in the plasma sheet and acceleration processes related to charged particle injections/dipolarizations(e.g., Gabrielse et al. 2014; Malykhin et al. 2018) as well as associated loss mechanisms. The proton intensities are also enhanced during such injections.

In this work, we have tried to include a wide range of the history of the solar and geomagnetic parameters. However, they did not improve the model. The solar wind–magnetosphere energy coupling functions (e.g., Gonzalez et al. 1994; Milan et al. 2012; Wang et al. 2014) are not considered in this work and are beyond the scope of this paper.

## 6. Conclusions and Outlook

Using 17 yr of Cluster/RAPID observations, we have derived a machine-learning-based model for predicting proton intensities at energies from 28 to 962 keV in the 3D terrestrial magnetosphere between 6 and 22 $R_E$. As predictors, we used the location, solar, solar wind, and geomagnetic activity indices. The results demonstrate that the neural network's (MLP) prediction capabilities exceed the baselines based on the kNN and historical binning on average by ∼80% and ∼33%, respectively. The average correlation between the observed and predicted data is about 56% despite the complex dynamics of the energetic protons in the magnetosphere. The most important parameters for predicting proton intensities in our model are the ZGSE direction and the radial distance, both related to location. The most important predictor of solar, solar wind, and geomagnetic activity is the solar wind dynamic

pressure. The results are in general agreement with the study by Kronberg et al. (2020) on the characteristics of the SP contamination in the XMM-Newton telescope. The results can directly be applied in practice, e.g., to assess the contamination of X-ray telescopes as well as better determine the contamination risk for various future mission concept orbits.

Based on this experience, in our future study, we plan to derive a tailored machine-learning model for predicting the proton intensities at different energy channels along the XMM trajectory. In a future study, we want to use the same time interval and location in the magnetosphere to avoid possible biases. The aim will be to find which energies of proton intensities correlate most strongly with the SP count/rates. The results of the tailored model will be compared with the results of this model.

# Appendix

Figure 9 demonstrates the distributions of the number of samples for the predictors and the proton intensities at different energies (on the vertical axis) in a corresponding range of values (on the horizontal axis).

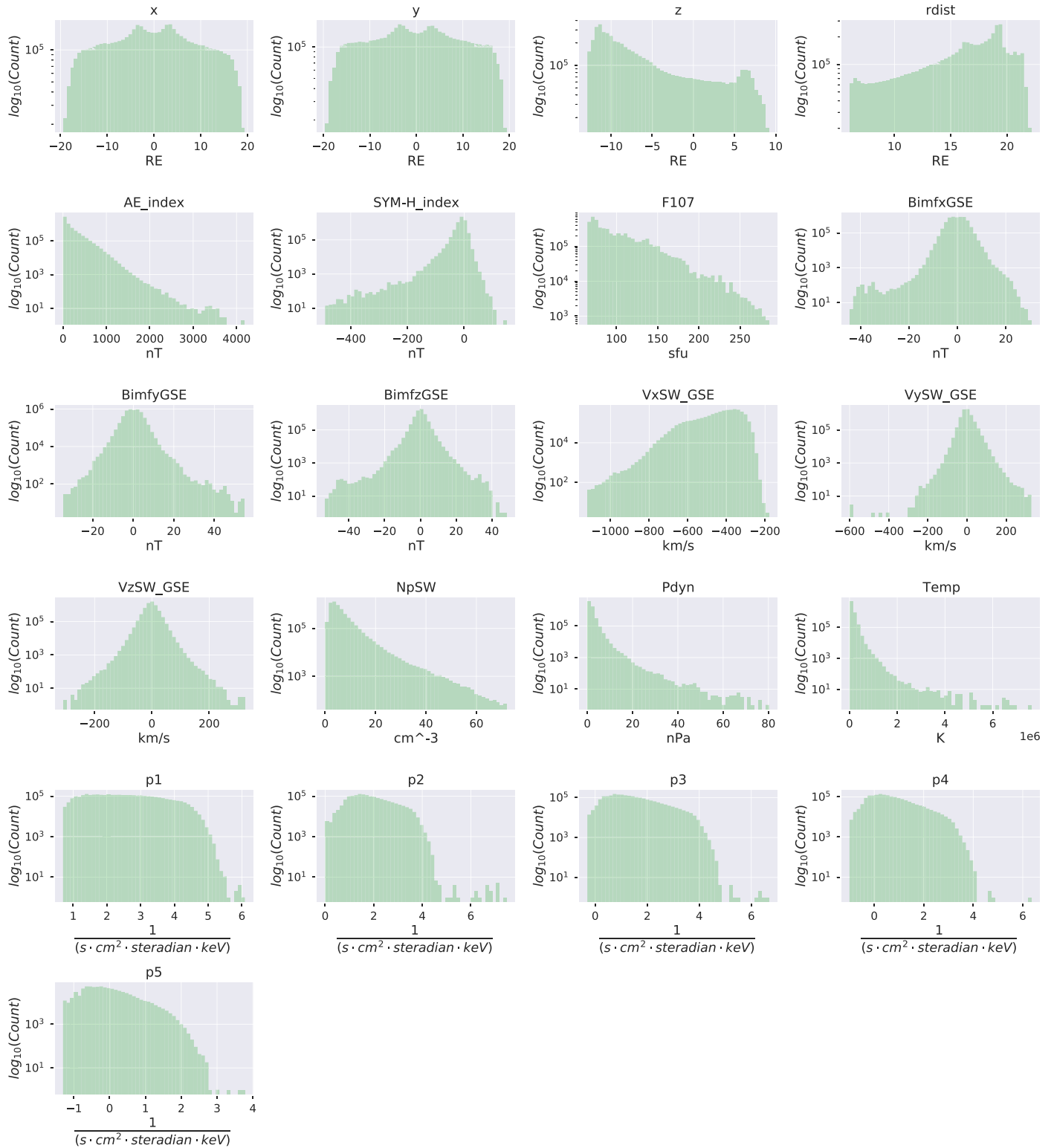Table 6 shows the search space for hyper-parameter optimization and the best values found for the MLP model.



**Figure 9.** Histogram of the number of samples of predictors and proton intensities used in the model. The intensities per channel are in $log_{10}$ space.

**Table 6**
Search Space for Hyper-parameter Optimization and the Best Values in the MLP Model

| Parameter | Search Space | Best Value |
|---|---|---|
| **Multi-layer Perceptron** | | |
| Number of Layers | (2, 5) | 2 |
| Hidden Dimensions | (32, 256) | (64, 32) |
| **Optimizer** | | |
| Batch Size | (32, 256) | 56 |
| Learning Rate | $(10^{-6}, 10^{-3})$ | 0.00046 |
| L2 Penalty | $(0, 10^{-2})$ | 0.00882 |

## ORCID iDs

Elena A. Kronberg ⓘ https://orcid.org/0000-0001-7741-682X
Tanveer Hannan ⓘ https://orcid.org/0000-0001-9957-9531
Jens Huthmacher ⓘ https://orcid.org/0000-0001-7223-4103
Marcus Münzer ⓘ https://orcid.org/0000-0002-8712-4035
Florian Peste ⓘ https://orcid.org/0000-0002-7070-1078
Ziyang Zhou ⓘ https://orcid.org/0000-0003-0154-6948
Max Berrendorf ⓘ https://orcid.org/0000-0001-9724-4009
Evgeniy Faerman ⓘ https://orcid.org/0000-0001-8841-5128
Fabio Gastaldello ⓘ https://orcid.org/0000-0002-9112-0184
Simona Ghizzardi ⓘ https://orcid.org/0000-0003-0879-7328
Philippe Escoubet ⓘ https://orcid.org/0000-0003-4475-6769
Stein Haaland ⓘ https://orcid.org/0000-0002-1241-7570
Artem Smirnov ⓘ https://orcid.org/0000-0003-3689-4336
Nithin Sivadas ⓘ https://orcid.org/0000-0003-4278-0482
Robert C. Allen ⓘ https://orcid.org/0000-0003-2079-5683
Andrea Tiengo ⓘ https://orcid.org/0000-0002-6038-1090
Raluca Ilie ⓘ https://orcid.org/0000-0002-7305-2579

## References

Altman, N. S. 1992, Am. Stat., 46, 175
Bergstra, J., & Bengio, Y. 2012, JMLR., 13, 281
Breiman, L., Friedman, J., Stone, C., & Olshen, R. 1984, Classification and Regression Trees (Oxford: Taylor & Francis)
Buitinck, L., Louppe, G., Blondel, M., et al. 2013, in ECML PKDD Workshop: Languages for Data Mining and Machine Learning (Berlin: Springer), 108
Cortes, C., & Vapnik, V. 1995, Mach. Learn., 20, 273
Daly, P. W., & Kronberg, E. A. 2010, in The Cluster Active Archive, Studying the Earth's Space Plasma Environment, ed. H. Laakso, M. Taylor, & C. P. Escoubet (Berlin: Springer), 145
De Luca, A., & Molendi, S. 2004, A&A, 419, 837
Escoubet, C. P., Fehringer, M., & Goldstein, M. 2001, AnGeo, 19, 1197
Facskó, G., Trotignon, J. G., Dandouras, I., Lucek, E. A., & Daly, P. W. 2010, AdSpR, 45, 541
Fioretti, V., Bulgarelli, A., Molendi, S., et al. 2018, ApJ, 867, 9
Fisher, A., Rudin, C., & Dominici, F. 2019, arXiv:1801.01489
Freund, Y., & Schapire, R. E. 1995, Lecture Notes in Computer Science (Berlin Heidelberg: Springer), 23
Friedman, J. H. 2001, AnSta, 29, 1189
Gabrielse, C., Angelopoulos, V., Runov, A., & Turner, D. L. 2014, JGRA, 119, 2512
Galton, F. 1886, JRAI, 15, 246
Geron, A. 2019, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (Sebastopol: O'Reilly Media, Inc.)
Geurts, P., Ernst, D., & Wehenkel, L. 2006, Mach. Learn., 63, 3
Ghizzardi, S., Marelli, M., Salvetti, D., et al. 2017, ExA, 44, 273
Gonzalez, W. D., Joselyn, J. A., Kamide, Y., et al. 1994, JGR, 99, 5771
Hoerl, A. E., & Kennard, R. W. 1970, Technometrics, 12, 55
Hunter, J. D. 2007, CSE, 9, 90
Iyemori, T., Araki, T., Kamei, T., & Takeda, M. 1992, Midlatitude Geomagnetic Indices ASY and SYM (provisional), Initial Report of Data Analysis Center for Geomagnetism and Space Magnetism, Kyoto Univ.
Jansen, F., Lumb, D., Altieri, B., et al. 2001, A&A, 365, L1
Ke, G., Meng, Q., Finley, T., et al. 2017, in 31st Conf. on Neural Information Processing Systems (NIPS 2017), ed. I. Guyon et al. (Red Hook, NY: Curran Associates, Inc.), 3146
King, J. H., & Papitashvili, N. E. 2005, JGRA, 110, A02104
Kronberg, E. A., Bučík, R., Haaland, S., et al. 2011, JGRA, 116, 2210
Kronberg, E. A., Gastaldello, F., Haaland, S., et al. 2020, ApJ, 903, 89
Kronberg, E. A., Grigorenko, E. E., Haaland, S. E., et al. 2015, JGRA, 120, 3415
Kronberg, E. A., Haaland, S. E., Daly, P. W., et al. 2012, JGRA, 117, 12208
Kronberg, E. A., Kis, A., Klecker, B., Daly, P. W., & Lucek, E. A. 2009, JGRA, 114, 3211
Kuntz, K. D., & Snowden, S. L. 2008, A&A, 478, 575
Leccardi, A., & Molendi, S. 2008, A&A, 486, 359
Lee, M. A. 1982, JGR, 87, 5063
Li, L., Jamieson, K., Rostamizadeh, A., et al. 2020, arXiv:1810.05934
Liaw, R., Liang, E., Nishihara, R., et al. 2018, arXiv:1807.05118
Lloyd, S. P. 1982, IEEE Trans. Inf. Theory, 28, 129
Lotti, S., Mineo, T., Jacquey, C., et al. 2018, ExA, 45, 411
Luo, H., Kronberg, E. A., Nykyri, K., et al. 2017, JGRA, 122, 5168
Malykhin, A. Y., Grigorenko, E. E., Kronberg, E. A., et al. 2018, AnGeo, 36, 741
McKinney, W. 2010, in Proc. of the 9th Python in Science Conf., ed. S. van der Walt & J. Millman (Austin, TX: SciPy), 56
Meng, C.-I., Lui, A. T. Y., Krimigis, S. M., Ismail, S., & Williams, D. J. 1981, JGRA, 86, 5682
Microsoft Corporation 2008, LightGBM 3.1.0.99 documentation, https://lightgbm.readthedocs.io/
Milan, S. E., Gosling, J. S., & Hubert, B. 2012, JGRA, 117, A03226
Nandra, K., Barret, D., Barcons, X., et al. 2013, arXiv:1306.2307
Nosé, M., Sugiura, M., Kamei, T., & Iyemori, T. 2017, Geomagnetic AE index, World Data Center for Geomagnetism, Kyoto, doi:10.17593/15031-54800
Nykyri, K., Otto, A., Adamson, E., Dougal, E., & Mumme, J. 2011, JGRA, 116, A03228
Nykyri, K., Otto, A., Adamson, E., Kronberg, E., & Daly, P. 2012, JASTP, 87, 70
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Raab, W., Branduardi-Raymont, G., Wang, C., et al. 2016, Proc. SPIE, 9905, 990502
Rosenblatt, F. 1958, Psychol. Rev., 65, 386
Rosenqvist, L., Hilgers, A., Daly, E., & Evans, H. 2002, JASTP, 64, 721
Santosa, F., & Symes, W. W. 1986, SIAM J. Sci. Statist. Comput., 7, 1307
Savin, S., Amata, E., Budaev, V., et al. 2014, JETPL, 99, 16
Smirnov, A. G., Kronberg, E. A., Latallerie, F., et al. 2019, SpWea, 17, 553
Tapping, K. F. 2013, SpWea, 11, 394
Tibshirani, R., Johnstone, I., Hastie, T., & Efron, B. 2004, AnSta, 32, 407
van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, CSE, 13, 22
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261
Wang, C., Han, J. P., Li, H., Peng, Z., & Richardson, J. D. 2014, JGRA, 119, 6199
Weisskopf, M. C., Brinkman, B., Canizares, C., et al. 2002, PASP, 114, 1
Wilken, B., Daly, P. W., Mall, U., et al. 2001, AnGeo, 19, 1355