Prediction of Plasma Pressure in the Outer Part of the Inner Magnetosphere using Machine Learning

Songyan Li[1], Elena A. Kronberg[2], Christopher G. Mouikis[3], Hao Luo[1], Yasong S Ge[4], and Aimin Du[5]

[1]Institute of Geology and Geophysics, Chinese Academy of Sciences
[2]Ludwig Maximilian University of Munich
[3]University of New Hampshire
[4]Institute of Geology and Geophysics
[5]Institue of Geology and Geophysics, Chinese Academy of Sciences

December 16, 2022

Abstract

The information on plasma pressures in the outer part of the inner magnetosphere is important for simulations of the inner magnetosphere and the better understanding of its dynamics. Based on 17-year observations from both CIS and RAPID instruments onboard the Cluster mission, we used machine-learning-based models to predict proton plasma pressures at energies from ˜40eV to 4MeV in the outer part of the inner magnetosphere (L*=5-9). The location in the magnetosphere, and parameters of solar, solar wind, and geomagnetic activity from the OMNI database are used as predictors. We trained several different machine-learning-based models and compared their performances with observations. The results demonstrate that the Extra-Trees Regressor has the best predicting performance. The Spearman correlation between the observations and predictions by the model data is about 68%. The most important parameter for predicting proton pressures in our model is the L* value, which is related to the location. The most important predictor of solar and geomagnetic activity is the solar wind dynamic pressure. Based on the observations and predictions by our model, we find that no matter under quiet or disturbed geomagnetic conditions, both the dusk-dawn asymmetry at the dayside with higher pressures at the duskside and the day-night asymmetry with higher pressures at the nightside occur. Our results have direct practical applications, for instance, inputs for simulations of the inner magnetosphere or the reconstruction of the 3-D magnetospheric electric current system based on the magnetostatic equilibrium, and can also provide valuable guidance to the space weather forecast.

Prediction of Plasma Pressure in the Outer Part of the Inner Magnetosphere using Machine Learning

S. Y. Li[1,2,3,4], E. A. Kronberg[4,*], C. G. Mouikis[5], H. Luo[1,2,3,], Y. S. Ge[1,2,3,], A. M. Du[1,2,3,]

[1]CAS Engineering Laboratory for Deep Resources Equipment and Technology, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China

[2]Innovation Academy for Earth Science, CAS, Beijing 100029, China

[3]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[4]Department of Earth and Environmental Sciences (Geophysics), Ludwig Maximilian University of Munich (LMU) Munich, Theresienstr. 41, Munich, 80333, Germany

[5]Department of Physics and Space Science Center, University of New Hampshire, Durham, New Hampshire, USA

*Corresponding Author: kronberg@geophysik.uni-muenchen.de

Key Points:

1. A machine learning model is created to predict 3-D distribution of proton plasma pressures at L*=5-9 for energies ~40eV-4MeV
2. Our model based on Extra-Trees Regressor reproduces well the global distributions as well as the pressure along a spacecraft trajectory
3. The results of our model are helpful for the interpretation of the plasma pressure in the outer part of the magnetosphere

Plain Language Summary

The distribution of the plasma pressures in the magnetosphere is a key parameter for the assessment of the magnetostatic equilibrium, the dynamics of geomagnetic storms, and the magnetospheric electric current system. In addition, the outer part of the inner magnetosphere (L^*=5-9) is often used as the boundary in the inner magnetosphere simulations, where the initial composition is specified. Thus, the distribution of the plasma pressure at L^*=5-9 is essential for the simulations of the inner magnetosphere and understanding of the underlying magnetospheric dynamic processes. Although, there are many previous studies on the distribution of plasma pressures, building a model to predict the 3-D distribution of plasma pressures remains challenging. Based on 17 years of data from both CIS and RAPID instruments onboard the Cluster spacecraft mission, a machine-learning-based model for predicting proton pressures at energies from ~40eV to 4MeV in the outer part of the inner magnetosphere (L^*=5-9) is built. We set up the 3-D model for the prediction of the proton pressures depending on the location, solar, solar wind, and geomagnetic activity indices. The model gives reliable predictions and can be used for the interpretation of the dynamics of the inner magnetosphere under different geomagnetic conditions which can also provide valuable guidance to the space weather (such as magnetic storms) forecast.

Abstract

The information on plasma pressures in the outer part of the inner magnetosphere is important for simulations of the inner magnetosphere and the better understanding of its dynamics. Based on 17-year observations from both CIS and RAPID instruments onboard the Cluster mission, we used machine-learning-based models to predict proton plasma pressures at energies from ~40eV to 4MeV in the outer part of the inner magnetosphere (L^*=5-9). The location in the magnetosphere, and parameters of solar, solar wind, and geomagnetic activity from the OMNI database are used as predictors. We trained several different machine-learning-based models and compared their performances with observations. The results demonstrate that the Extra-Trees Regressor has the best predicting performance. The Spearman correlation between the observations and predictions by the model data is about 68%. The most important parameter for predicting proton pressures in our model is the L^* value, which is related to the location and distance. The most important predictor of solar, solar wind, and geomagnetic activity is the

solar wind dynamic pressure. Based on the observations and predictions by our model, we find that no matter under quiet or disturbed geomagnetic conditions, both the dusk-dawn asymmetry at the dayside with higher pressures at the duskside and the day-night asymmetry with higher pressures at the nightside occur. Our results have direct practical applications, for instance, inputs for simulations of the inner magnetosphere or the reconstruction of the 3-D magnetospheric electric current system based on the magnetostatic equilibrium, and can also provide valuable guidance to the space weather forecast.

1. Introduction

In the inner magnetosphere, the plasma pressure plays a key role in the understanding of the main magnetospheric dynamic processes. The knowledge about distributions of the plasma pressures under different geomagnetic conditions is necessary for explaining how the Earth's magnetosphere reaches the magnetostatic equilibrium (the plasma pressure gradient is compensated by Ampere's force) and what specific conditions are necessary to maintain it (Antonova, 2004; Stepanova et al., 2019). In addition, one of the key parameters for understanding the evolution of geomagnetic storms and substorms is the plasma pressure distribution in the inner magnetosphere (Kronberg et al., 2017; Stepanova et al., 2008). The increase of the inner magnetospheric plasma pressure is one of the main features of magnetic storms (Stepanova et al., 2019).

The distribution of plasma pressures in the inner magnetosphere has been studied extensively during last decades. Based on the measurements from the high-altitude AMPTE/CCE satellite, Lui and Hamilton (1992) obtained average radial profiles of plasma pressures from a case study during geomagnetically quiet conditions. These profiles showed a peak generally at L = 3 to 4 and decreased from L = 4 to L = 9 rather monotonically. Using data from the same satellite, De Michelis et al. (2013) presented the statistical study of plasma pressure profiles which were averaged over more than 2 years data. The low-activity pressure profile gave the same general features as the profiles in the case study published by Lui and Hamilton (1992). The disturbed pressure profile had a peak at a higher L value (L~4.5) and decreased from L = 5 to L = 8, also rather monotonically. The equatorial plasma pressure distribution can be obtained from low-altitude measurements under the assumption that the plasma pressure is conserved along a magnetic field line (Wing & Newell, 1998). Based on the energetic neutral

atom (ENA) images obtained by HENA onboard the IMAGE spacecraft, Brandt et al. (2004) inferred the evolution of the global plasma pressure distribution during storms, showing that there was a peak of the proton pressure located around the midnight. Similarly, Lui (2003) found that the proton pressures were generally higher in the dusk-midnight sector than in the post-midnight sector under disturbed geomagnetic conditions within the L shells from 2 to 9 in the equatorial region. Using the data from both low-orbiting (DMSP 16–18 spacecraft, satellites NOAA 15–19, and METOP 1–2 satellites) and high-orbiting satellites (the THEMIS and Van Allen Probes), Stepanova et al. (2019) performed a multisatellite analysis of the variation of plasma pressures near the equatorial plane between 7 to 13 R_E during a strong geomagnetic storm. They also found that the plasma pressure inside the magnetosphere is mainly controlled by the solar wind dynamic pressure.

Among these previous studies, most of them focus on the 2-D equatorial plasma pressure distribution (Antonova et al., 2014; Lui, 2003; Stepanova et al., 2019; Wing & Newell, 1998). The 3-D plasma pressure distribution in the inner magnetosphere is relatively unknown. This kind of 3-D distribution is not only helpful to understand the dynamics of the inner magnetosphere, but also important for the simulations of the inner magnetosphere. For example, it can be used to deduce the distribution of the temperature which is an important input for some simulations models (such as Hot Electron Ion Drift Integrator (HEIDI) model (Ilie et al., 2012)). In addition, using the steady−state force balance equation $\nabla P = j \times B$ (Sergeev et al., 1994; Stephens et al., 2013) allows one to reconstruct the 3-D electric current system with the 3-D plasma pressure distribution.

In this study, we derive a predictive model for the proton pressures at energies from ~40eV to 4MeV in the outer part of the inner magnetosphere (L^*=5-9). For this energy range, we combined the data from both CIS and RAPID instruments onboard Cluster. This L^* range is selected because it is the region that is often used as the boundary in the inner magnetosphere simulations, where the initial composition is specified (Kistler & Mouikis, 2016). Instrumentally, we restricted the minimum distance to L^*=5 to reduce the contamination of the results by energetic electrons from the outer radiation belt. To enable modeling of the complex non-linear multidimensional dependencies, we trained several different machine-learning-based models and compared their performances with the observations. Moreover, by

using a machine learning method we can utilize the full range of solar or geomagnetic parameters as inputs to infer and analyze the plasma pressure distribution instead of only considering a few of them as most previous studies did. This helps to increase the performance of the predictions.

To summarize, our study aims are to (1) test the capability of different machine learning algorithms and to present the best one for the prediction of the proton plasma pressures in the outer part of the inner magnetosphere; (2) reveal which parameters are the most important for the prediction of proton plasma pressures; (3) compare and analyze the prediction of the proton plasma pressure distributions under different geomagnetic conditions and (4) help future studies that require a proton plasma pressure model. The remainder of this paper is organized as follows. In section 2 we describe the observations and data analysis. Section 3 is concerned with the methodology used for this study. Section 4 presents the results. The discussions and conclusions are drawn in the last two sections 5 and 6.

2. Observations and Data Set

In this section, we first introduce the data and the method we used for calculating proton plasma pressures. Then, the predictors, also called features, (the variables that are potentially capable to predict the proton pressures) are also discussed.

2.1 Instrumentation and Data

The Cluster mission consists of four identical spacecrafts, each carrying 11 instruments. The satellites were launched in two pairs in late 2000, and after a 6-months commissioning phase, the mission moved into an operational phase in February 2001 (Laakso et al., 2010). For the first ~6 years of the mission, the spacecrafts were placed into a highly elliptical, polar orbits apogee at 19.6 Re and perigee at 4 Re (Escoubet et al., 1997). The orbit has evolved over time, passing through the inner magnetosphere closer to the equator. It also covers the full range of local times over the course of a year (Kistler & Mouikis, 2016). Thus, the data from the Cluster mission is proper for the study of 3-D distribution of proton pressures in the inner magnetosphere. We used the proton observations from the spacecraft (SC) 4 (Tango) since for this study it is necessary that particle instruments have been operating nearly continuously from 2001 through the present day. The L^* range we chose to study is L^*=5-9 as we introduced above.

Based on the observations by the Cluster Ion Spectrometry (CIS) using the time-of-flight ion Composition Distribution Function (CODIF) sensor (Reme et al., 2001), the low-energy component of the proton pressure (P_{CIS}) with the energy range of ~40 eV to 40 keV is calculated using the formula

$$P_{CIS}[nPa] = 1.381 \cdot 10^{-2} n[cm^{-3}] T[MK],$$

where n is the proton density, T is the proton temperature. The proton densities and temperatures can be found at CSA under the product C4_CP_CIS-CODIF_HS_H1_MOMENTS.

Based on the observations by the Research with Adaptive Particle Imaging Detector (RAPID) (Wilken et al., 2001), the high-energy part of the proton pressure (P_{RAPID}) with the energy range up to 4MeV is calculated using the formula (Daly & Kronberg, 2018; Kronberg et al., 2017):

$$P_{RAPID}[nPa] = 4\pi \frac{2}{3} 0.518 \cdot 10^{-8} \sqrt{m[amu]} \sum_{E} \Delta E[keV] \sqrt{E[keV]}\, j[cm^{-2} sr^{-1} s^{-1} keV^{-1}],$$

where m is the proton mass in atomic mass units (amu), j is the omnidirectional energetic proton intensity, ΔE is the width of the energy channel, and E is the effective energy. The geometric mean is used as an approximation for the effective energy (Kronberg & Daly, 2013). The omnidirectional energetic proton intensities can be found at CSA under the product Proton_Dif_flux__C4_CP_RAP_HSPCT. The first proton RAPID energy channel at 27.7–64.4 keV overlaps with the last CIS energy channel. Thus, in order to obtain a continuous spectrum, the first RAPID channel was truncated according using the method provided by Kronberg et al. (2022). The CIS and the RAPID instruments are well cross-calibrated for protons (Kronberg et al., 2010, 2022).

The steps of data processing are as follows. First, the proton data from both CIS and RAPID with original 4-second resolution was averaged over 1 minute in order to be consistent with the predictors related to the solar and geomagnetic activity which have the highest resolution of 1 minute. Then, the outliers were eliminated. The data points with calculated $P_{CIS} = 0$ were removed because both proton densities and temperatures were 0 at these points. When calculated $P_{CIS}$was above 100nPa, the data points were also removed. Because based on previous studies (De Michelis et al., 2013; Kronberg et al., 2017; Lui, 2003), proton pressures have never been over 100nPa even under disturbed geomagnetic

conditions. The large P_{CIS} at these points may be caused by the large proton densities (>100 cm^{-3}) due to the background contamination of CODIF. When calculating P_{RAPID}, we had to remove the data points with the flux of the second energy channel (75.3-92.2 keV) being 0. The slope of the energy spectra to derive the fluxes in the first truncated channel cannot be calculated in this case, leading to the inaccurate P_{RAPID}. Next, we added the P_{CIS} and P_{RAPID} with the same timestamp to get the total proton pressure. Finally, we chose the points with the position range of L^*=5-9 for our study. We also removed the points with the total proton pressures less than 0.1nPa because these values are not typical at these distances on the closed magnetic field lines. Most previous work showed that the values of the proton pressures were mostly above 0.1nPa in the inner magnetosphere (Kronberg et al., 2017; Lui, 2003). In addition, there were only 1580 points (0.5% of the total points) less than 0.1nPa. We obtained better model performance when dropping the values less than 0.1nPa. In this case, the model was more focused on predicting values above 0.1nPa, namely the typical values of proton pressures in the inner magnetosphere.

2. 2 Predictors

In this subsection, we divide predictors into two groups for description: related to the location in the space and related to the solar, solar wind, and geomagnetic activity. All the predictors are listed in **Table 1**.

2.2.1 Location in the Space

The predictors related to the location in the space include: L^* value, magnetic local time (MLT), and the position of Cluster in the Geocentric Solar Ecliptic (GSE) coordinate system (x_gse, y_gse and z_gse). The L^* values were taken from the Lstar_value__C4_CP_AUX_LSTAR dataset. MLTs can be found under the product Mag_Local_time__C4_JP_AUX_PMP. The sc_r_xyz_gse__C4_CP_AUX_POSGSE_1M dataset is the source of the position data. The distributions of the proton plasma pressures and the numbers of their samples in the GSE system are shown in **Figure 1**. The distributions of the proton plasma pressures and the number of their samples in the L^*-MLT coordinate are shown in **Figure 2**.

From **Figures 1d-f**, we can see that the number of samples is larger on the dayside and especially

at lower L^* shells. As we mentioned above, our L^* values are from "Lstar" product (A. G. Smirnov et al., 2020) which is calculated with Tsyganenko-89 magnetospheric model, T89 (Tsyganenko, 1989). Thus, one can understand that the Cluster trajectories can cover more samples in the dayside than in the nightside at the same L^*-shell range. This phenomenon can also be seen in **Figure 2b**, the numbers of samples in the L^*-MLT coordinate. However, this has no effects on the final results because the value in each bin is the median of at least hundreds of samples, which is enough to reduce the error and represent the median feature of the bin.

Figures 1a and **1b** show that the proton pressures are higher at the nightside than that at the dayside in the XY and XZ planes in the GSE coordinate system. Similarly, the same result can be observed in the L^*-MLT coordinate at L^*-shell > 5 in **Figure 2a**. **Figure 1a** also shows the dusk-dawn asymmetry at the dayside with higher proton pressures at the dusk side. The same higher pressures at the dusk side (+y side) are visible in **Figure 1c** (YZ plane). Likewise, we can note this dusk-dawn asymmetry at the dayside at lower L^* shells in **Figure 2a.** The reasons for these asymmetries will be discussed in **Section 5**.

In **Figures 3a-e**, the relations of mean proton pressures versus the predictors related to the location in the space are shown. **Figure 3a** shows a strong linear decrease of the proton plasma pressure with the L^* shell which is consistent with previous results as we introduced in **Section 1.** That is, no matter under quiet conditions (Lui & Hamilton, 1992) or disturbed conditions (De Michelis et al., 2013), the proton plasma pressure is monotonically decreasing with L^* shell when L^*=5-9. In **Figure 3b**, there is a peak of the proton pressure around midnight (0-2MLT) and the minimum of proton pressures is shown around 9 MLT. The proton pressure displays a peak at ~-6R_E and roughly linear decrease with XGSE coordinate in the distances between -6 R_E and 9 R_E in **Figure 3c**. This is also consistent with the day-night asymmetry with higher proton pressures at the nightside (-x side) in **Figures 1a** and **1b**. **Figure 3d** shows that the maximum of the proton pressure is around ±5 RE in YGSE direction. This YGSE dependency resembles that observed for the proton intensities in the near-Earth space in Kronberg et al. (2021). In **Figure 3e**, we can take ZGSE=-2R_E as the symmetry axis, the proton pressures on both sides decrease almost linearly with the direction away from the symmetry axis.

2.2.2 Solar, Solar Wind and Geomagnetic Activity

The predictors are also related to the observations of solar, solar wind, and geomagnetic parameters from the OMNI database (King & Papitashvili, 2005). The solar wind parameters we used include: the proton density, NpSW [cm^{-3}]; components of the velocity (VSW) in the GSE coordinates, VxSW_GSE, VySW_GSE and VzSW_GSE [$km \cdot s^{-1}$]; the proton temperature, Temp [K] and components of the IMF in the GSE coordinates, BimfxGSE, BimfyGSE and BimfzGSE [nT]; and the dynamic pressure, Pdyn [nPa]. The proton pressures increase with the solar wind velocity in the anti-sunward direction, V_x, in **Figure 3f.** The V_y and V_z components are associated with increase of the proton pressures when they deviate from the Earth-Sun direction (V_x) within a certain range (<±100 km/s), as shown in **Figure 3g**. When any component of the interplanetary magnetic field (IMF) becomes stronger, no matter in positive or negative directions, it may lead to the increased proton pressures (see **Figure 3h**). These dependencies of the proton pressures on the solar wind velocity components and the IMF components also resemble those observed for the proton intensities in the near-Earth space in Kronberg et al. (2021). In order to show the relationships more clearly, we plot the figures in the logarithmic scale in **Figures 3i-k**. In **Figure 3i**, when the solar wind density is greater than 2 cm^{-3}, the proton pressures generally increase with it. There is an approximate linear relationship between the proton pressures and the logarithm of the solar wind dynamic pressures when the solar wind dynamic pressure is greater than 1nPa (see **Figure 3j**). In **Figure 3k**, a trend of increase of the proton pressures with the solar wind temperature is visible. The 10.7 cm solar radio flux ($F_{10.7}$) is one of the most widely used indices of solar irradiance (Tapping, 2013). Kistler and Mouikis (2016) showed that $F_{10.7}$ have an impact on the proton flux at L=6-7. In **Figure 3l**, the solar irradiance is non-linearly related to the proton pressure, but a general trend of decrease of the proton pressures with the F10.7 is visible.

For the parameters related to the geomagnetic activity, we used AE index and SYM-H index. The auroral electrojet index (AE index) provides a global, quantitative measure of auroral zone magnetic activity within the auroral oval. The proton pressures are related roughly linear up to ~ 600nT with the logarithm of the AE index in **Figure 3m**. SYM-H index describes symmetric horizontal component disturbances of the geomagnetic field at the equatorial regions (Iyemori et al., 2010). SYM-H index,

shows non-linear relation with proton pressures, see **Figure 3n**. The histograms of the number of samples of all these predictors are shown in **Figure 12** in **Appendix**.

2.2.3 Cross-correlations between Proton Pressures and Predictors

In **Figure 4,** we show the Pearson linear correlations between proton pressures and the predictors. This measurement can only reflect a linear correlation of variables, and ignore other types of correlations. The range of this correlation value is from -1 to 1. Values close to -1/1 mean perfect linear anticorrelation/correlation and values equal to 0 mean there is no linear dependency between the variables. The proton pressures are well anticorrelated with the L^* shell (-0.51), in agreement with **Figure 3a**. For the XGSE and ZGSE location of observation, they also show some anticorrelation with the proton pressures, -0.21 and -0.14, respectively. From the OMNI parameters, the proton pressures are best linearly correlated with the solar wind dynamic pressure, 0.23. The AE-index also shows some correlation with proton pressures (0.16), the same as the result in **Figure 3**.

3. Methodology

3.1 Data Split

After the data processing as we mentioned in **Section 2.1**, the full dataset we used comprises in total 336481 measurements from 2001-02-04 12:31:00 UT to 2018-02-18 00:02:00 UT. We split the dataset into a training set (80%) and a test set (20%). To prevent test leakage, we split the data by a time point with the original order preserved (Camporeale, 2019; Kronberg et al., 2021). The test set is only for the testing of the model. After the model training has been completed, no further changes to the model can be made. We utilize the training set to train and optimize the model hyperparameters. The sizes and periods of data subsets after splitting are listed in **Table 2**.

Machine Learning algorithms don't perform well when the input numerical features have very different scales. Thus, we need to normalize the features in order to get all the features to have the same scale (Géron, 2019). We normalized the features by QuantileTransformer in sklearn (Pedregosa et al., 2011). QuantileTransformer provides a non-parametric transformation to map the data to a uniform distribution with values between 0 and 1. This transformation smooths out unusual distributions and is

less influenced by outliers than other methods (Pedregosa et al., 2011).

3.2 Machine Learning Models for Proton Pressures

Our study is one of the typical supervised learning tasks, called regression. We have applied various kinds of regression ML models in order to select the best one based on their validation performance. We note that most of the relations between the proton pressures and the predictors are not perfectly linear, as shown in **Figure 3.** In addition, the ensemble of the predictions of a group of predictors (such as regressors) will often give better predictions than with the best individual predictor (Géron, 2019). Thus, we not only tried linear regression models, but also ensemble regression models.

We have examined the following linear models in sklearn.linear_model and sklearn.svm (Pedregosa et al., 2011): (1) Ridge Regression, namely linear least squares with l2 regularization (Hoerl & Kennard, 1970); (2) Least Angle Regression (LARS) (Efron et al., 2004); (3) Linear Support Vector Regression (LinearSVR) (Cortes & Vapnik, 1995).

We also consider the Decision Trees Regression (Breiman et al., 1984) in sklearn.tree and the tree-based ensemble models: (1) Random Forest Regression (Ho, 1995); (2) Extra Trees Regression, namely extremely randomized trees (Geurts et al., 2006); (3) AdaBoost Regression (Freund & Schapire, 1997); (4) Gradient Boosting Regression (Friedman, 2001); (5) Histogram-Based Gradient Boosting Regression ((1)-(5) are all from sklearn.ensemble); (6) Light Gradient Boosting Machines (LGBM) (Ke et al., 2017) in LightGBM library.

In order to evaluate different models' performances, we focus on the Spearman correlation. The Spearman correlations between the model results and the observations are listed in **Table 3.** Pearson correlation only assesses linear relationships as we discussed in **Figure 4**, while Spearman correlation assesses monotonic relationships (whether linear or not). The values of the Spearman correlation vary between -1 and 1. Correlations of -1 or +1 imply an exact monotonic relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases. Values close to 0 means no monotonically correlation. In **Table 3**, we can note that the Extra-Trees Regressor has shown the best predicting performance on the both sets. Although the Decision Tree Regressor also

has a perfect performance on the train/validation set, it seems to be more inclined to the overfitting (the difference between the scores for the train/validation set and test set are larger). In addition, note that a gap between model performance on training and test data is often observed for complex models (Kronberg et al., 2021). Extra-Trees Regressor fits a large amount of randomized decision trees on the training dataset and uses the mean to improve the predictive accuracy and control overfitting. It has two main differences with other tree-based ensemble methods: (1) it splits nodes by choosing cut-points fully at random. That is, besides searching for the best feature among a random subset of features, like the regular Random Forests, it also utilizes random thresholds for each feature rather than searching for the best possible thresholds. (2) it uses the whole learning sample (rather than a bootstrap replica) to grow the trees (Géron, 2019; Geurts et al., 2006). These characteristics can result a lower variance and a faster training speed (compared with regular Random Forests). Thus, we decided to use Extra-Trees Regressor.

3.3 Training the selected model

We trained the model using the K-Fold cross-validation (CV) function (from sklearn.model_selection.KFold). This method is widely used in previous work (e.g., Kronberg et al., 2020; A. Smirnov et al., 2020). Our training data are divided into K subsets (folds) which are roughly the same size. In our case, K =5. Then each fold is used once as a validation while the 4 remaining folds form the train set. In this way, splitting process is repeated 5 times and results in five arrays of evaluation scores. Cross-validation allows one to get not only an estimate of the performance of the model, but also a measure of how precise this estimate is (Géron, 2019).

In order to determine the best hyperparameters, the parameters are optimized by grid-search over a parameter grid, using GridSearchCV (from sklearn.model_selection.GridSearchCV). To evaluate the performance of the training and validation during the cross-validation for different parameters, we use four assessment metrics: Spearman correlation, mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2). The best scores for MSE and MSE are close to 0. R^2 is a number between 0 and 1 that measures how well a statistical model predicts an outcome. That is, comparing the case of using the model for prediction with the case of only using the mean prediction, to see how much the performance of the model has been improved. In the perfect case, R^2 is equal to 1. The resulting

hyperparameters values, as well as their search ranges, are given in **Table 4**. The performances of the model for the train/validation data set are mostly consistent between different metrics. n_estimators controls the number of trees in the forest. It will be underfitting when n_estimators is too small, while it will be overfitting when n_estimators is too large. max_depth is the maximum depth of the tree. If max_depth=None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. min_samples_split is the minimum number of samples required to split an internal node. We use the default value 2 for min_samples_split. Another important parameter we optimized is the minimum number of data points in leaf (min_data_in_leaf), which has a regularization effect and stops the model from learning the noise.

4. Results

4.1 Test the model

The final scores are the average performances of the model for the train/validation data set, see **Table 5**. The values of the Spearman correlation coefficient are very close for the test data (0.68) and the average validation (0.71). The mean squared (MSE) and absolute errors (MAE) also yield almost identical values for validation and test sets. This means that our model is not overfitting and successfully learns relationships between the input parameters and the resulting proton pressures and generalizes well onto the unseen data.

The Spearman correlation between the observed and predicted data is about 68% for the test set. This value is reasonable considering the complex dynamics of the energetic protons in the inner magnetosphere. In addition, when we use the Spearman correlation to evaluate our model, there is a null hypothesis states that the predictions are uncorrelated to the observations (evaluated by p value). We obtain p=0. In other words, we can reject the null hypothesis, namely the model predictions are correlated to the observations. Thus, our model results are reliable and can learn the overall trend in the proton pressures.

4.2 Visualized Results

Figure 5a shows the distribution of the observed proton plasma pressures versus the predicted values from the training set, while **Figure 6a** represents the test set distribution. Observed and predicted

data for the training and test data sets agree relatively well. The diagonal shows the one−to−one ratio between the observed and predicted pressures. The data is mainly concentrated along the black dashed line, corresponding to a good correlation. The histograms in **Figures 5a** and **6a** represent the predicted or observed data points that fall into each corresponding bin. **Figures 5b** and **6b** provides the histogram of model residuals. From these figures, we can note that for both the training set and the test set, our model has very low bias. Most of the model residuals are within the range of ±0.5, namely the ratios of observations and predictions are valued in ~0.3-~3 ($10^{-0.5}$-$10^{0.5}$). Therefore, we can conclude that our final model predicts the proton pressures at L^*=5-9 well since it has low bias and can capture the general trends represented in the data.

In **Figure 7**, we show a qualitative example of the model's predictions within the 6-hour time interval on 2017-07-04 (in the test set) under quiet geomagnetic conditions (SYM_H>-1). The model almost predicts the same proton plasma pressures with the observations in **Figure 7c**. **Figure 8** shows another example on 2017-09-28 (in the test set) which demonstrates the model performance during the main phase of a magnetic storm with the SYM-H index dropping down to ~-68nT. We can note in the panel c that the predictions are almost always lower than the observations under disturbed geomagnetic conditions. This is because the main phases of the magnetic storms are rather rare events in our dataset. Our model is not developed specifically for the prediction of the proton pressures under disturbed geomagnetic conditions. The ML model of soft proton intensities by Kronberg et al. (2021) also has better prediction efficiency under quiet geomagnetic conditions.

We also plot the distributions of the predicted proton plasma pressures in L*-MLT coordinates under quiet (**Figure 7d**) and disturbed geomagnetic conditions (**Figure 8d**). The input predictors are the median values of the parameters over the time period (except the location parameters: X/Y/Z_GSE, L* value and MLT) in the purple region in **Figures 7** and **8**. The details of the input predictors are listed in **Table 6** in the **Appendix**. For calculations of L* values, it is necessary to specify the satellite position, magnetic field model and geomagnetic conditions (by 'get_Lstar' function in IRBEM library). The initial position range we give in each direction is [-11, 11] R_E in the GSE system, which is consistent with the X/Y/Z_GSE range of our observation dataset. Tsyganenko-89 magnetospheric model, T89 (Tsyganenko,

1989) with Kp index as an input is employed, as Smirnov et al. (2020) did. The Kp index for quiet geomagnetic times is set as 0, while for disturbed geomagnetic times is set as 4. The time moments indicated by the red dotted lines in **Figures 7** and **8** are specified for the calculations of MLTs. All of the calculations are performed using the spacepy.irbempy library ('get_Lstar' function). Finally, we select the data points that fit the range of L*=5-9 from the output results for plotting **Figures 7d** and **8d**.

We can note that there are no results when L*>~8 under disturbed geomagnetic conditions in **Figure 8d**. L* is the property of a stably trapped particle. A pseudo-trapped particle (particles that will leave the magnetosphere before completing a 180° drift) that drifts into the magnetopause (magnetopause shadowing) or into the tail (tail-shadowing) does not have an L*-value (Roederer, 1967; Roederer & Lejosne, 2018). The particles are easier become pseudo-trapped particles and to be lost during disturbed times (Roederer & Lejosne, 2018) since the magnetosphere is compressed based on T89 model (Tsyganenko, 1989).

From the predictions in **Figure 7d** and **8d**, we can note the dusk-dawn asymmetry at the dayside with higher proton pressures at the dusk side and the day-night asymmetry with higher proton pressures at the night side**.** In regard of the day-night asymmetry under quiet geomagnetic conditions in **Figure 7d** we consider the higher L^* shells (L^*>6). In addition, we can note that the proton pressures at the nightside under disturbed geomagnetic conditions seem to be higher than that under quiet geomagnetic conditions. The proton pressures at the afternoon sector (12-18MLT) under quiet geomagnetic conditions seem to be higher than that under disturbed geomagnetic conditions. A quantitative analysis of this phenomenon and the reasons for these asymmetries will be discussed in **Section 5**.

The plotting process of **Figure 9b** is the same as that of **Figures 7d** and **8d**, except that the input predictors are the median values of the parameters over the whole dataset time. The details of the input predictors are listed in **Table 6** in the **Appendix**. The Kp index is set as 0. The **Figure 9a** and **Figure 2a** is the same figure, namely the distribution of the observed H^+ plasma pressures over the whole dataset time. In **Figure 9b**, we can note the dusk-dawn asymmetry at the dayside and the day-night asymmetry, just as the observations in **Figure 9a**. By comparing the results of observations (**Figure 9a**) and predictions by our model (**Figure 9b**) based on the whole dataset, we can conclude that our model can

reproduce the overall characteristics of the distributions of observed proton plasma pressures in the range of L*=5-9.

4.3 Feature Importance

One of the advantages of the tree-based machine learning models is that they make it easy to measure the relative importance of each feature (Géron, 2019). Scores are automatically computed for each feature after training. The values of all the feature importances sum to 1. This process is called feature importance. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature, also known as the Gini importance. The higher the value, the more important the feature is. **Figure 10** shows the feature importance for each input variable. The black horizontal lines represent confidence intervals at 95% confidence level. The parameters related to the location show significantly higher importance than parameters related to solar, solar wind, and geomagnetic activity. From those, on average, the strongest dependence is seen for L^* shell. This is also consistent with the results in **Figures 3a** and **4**. The least important location parameter is y_gse. From the other parameters, the solar wind dynamic pressure is the most important parameter for predicting the proton plasma pressures. Based on the observations of multiple satellites, Stepanova et al. (2019) also found that the plasma pressure inside the magnetosphere is mainly controlled by the solar wind dynamic pressure, which can be related to the pressure balance at the magnetospheric flanks.

5. Discussions

Based on the observations and predictions by our model under quiet (**Figure 7**) and disturbed (**Figure 8**) geomagnetic conditions, we note that no matter under quiet or disturbed geomagnetic conditions, both the dusk-dawn asymmetry at the dayside and the day-night asymmetry occur. The persistent dusk-dawn asymmetry with higher proton pressures at the duskside may be related to the dawn-dusk asymmetry of the proton distribution in the plasma sheet. Based on 7-years observations of energetic protons >274 keV by RAPID instrument, Kronberg et al. (2015b) showed the dawn-dusk asymmetries of proton intensities in the plasma sheet at near-Earth nightside under both quiet and disturbed geomagnetic conditions. They also explained two general effects which can lead to this kind of dawn-dusk asymmetry. In addition, other previous work also reported dawn-dusk asymmetries in the

plasma sheet of energetic particles intensities with different energy ranges (Meng et al., 1981; Sarafopoulos et al., 2001; Kistler & Mouikis, 2016).

The day-night asymmetry is easy to understand because ions are injected into the inner magnetosphere through the plasma sheet at the nightside, especially during the magnetic storms or substorms (Kistler et al., 1992). Gabrielse et al. (2014) indicated that injection occurrence rates increase with the geomagnetic activity. This may also be the reason for the higher proton pressures at the nightside under disturbed geomagnetic conditions as we shown above. For further quantitative analysis of this phenomenon, **Figure 11** was plotted to investigate the difference between the proton plasma pressure under disturbed and quiet geomagnetic conditions. The red colors are positive values which means that the proton pressures are higher under disturbed geomagnetic conditions than under quiet geomagnetic conditions, while the blue colors are the opposite. The difference was calculated by the predictions of our model (**Figures 7d** and **8d**). In **Figure 11**, we can note that the proton pressures at the nightside under disturbed geomagnetic conditions are clearly higher than that under quiet geomagnetic conditions. In addition, more red bins are seen at the lower L* shells (L*=5-6) than at the higher L* shells (L*>6) at the nightside, which means that there are higher increases in the plasma pressures at the lower L* shells (L*=5-6) during disturbed times. This is consistent with the results of Figure 6b in Gabrielse et al. (2014). Namely, injections more frequently reach lower L-shells with increased geomagnetic activity. This can also be the reason why the day-night asymmetry under quiet geomagnetic conditions in **Figure 7d** mainly concentrates on the higher L^* shells (L^*>6).

In addition, the proton pressures at 12-18MLT sector (afternoonside) under disturbed geomagnetic conditions are clearly lower than that under quiet geomagnetic conditions. This may be related to the outflow of energetic ions through the magnetopause in the dayside under disturbed geomagnetic conditions. Keika et al. (2005) showed that the outflowing energy flux is higher on the afternoonside than that on the morningside during the main phase of magnetic storms, which may lead to the lower proton pressures on the afternoonside under disturbed geomagnetic conditions. Estimation of a comparison between the losses at the magnetopause and the difference between the proton plasma pressure on the afternoonside under disturbed and quiet geomagnetic conditions requires further studies.

Thus, we can deduce that the patterns of the asymmetries may change with the geomagnetic conditions. However, a more detailed calculation of the asymmetry index (e.g., Luo et al.,2017) separately for the quiet and the disturbed time is beyond the scope of this paper and will be further studied in the future.

6. Conclusions

In this study, based on 17-year data from both CIS and RAPID instruments onboard the Cluster mission, we derive a machine-learning-based model for predicting proton pressures at energies from ~40eV to 4MeV at the outer part of the 3D inner magnetosphere (L^*=5-9). The results demonstrate that the Extra-Trees Regressor shows the best predicting performance. The Spearman correlation between the observed and predicted data is about 68% despite the complex dynamics of the energetic protons in the magnetosphere. The most important parameter for predicting proton pressures in our model is the L^* shell, related to the location. The most important predictor of solar, solar wind, and geomagnetic activity is the solar wind dynamic pressure. The model results are in general agreement with the previous studies (De Michelis et al., 2013; Lui & Hamilton, 1992; Stepanova et al., 2019). In addition, we use the model prediction to compare and explain the distributions of the proton plasma pressures under different geomagnetic conditions. Moreover, as we discussed in the introduction, our results can be used in the simulations of the inner magnetosphere (e.g., HEIDI model) or reconstructing the 3-D electric current system. It can also provide valuable guidance to the space weather forecast.

Further directions for the present study include, first, incorporating oxygen ions data into the model in order to predict the complete 3D distribution of ion plasma pressures in the outer part of the inner magnetosphere. Second, a machine-learning-based model for predicting the 3-D ion pressures in the inner part of the inner magnetosphere (L^*=2-5). This aim can be achieved by using data from other missions, such as Van Allen Probes. The results of the model for the ion pressures in the inner part of the magnetosphere will be compared with the results of this model. In addition, we can combine these two models together to predict the 3-D ion pressures in the complete inner magnetosphere (L^*=2-9).

Acknowledgement

The authors are thankful to the Cluster Science Archive team (https://csa.esac.esa.int) for providing the

data. We acknowledge the use of NASA/GSFC's Space Physics Data Facility's OMNIWeb service and OMNI data. We acknowledge the use of the IRBEM library (V4.3), the latest version of which can be found at https://doi.org/10.5281/zenodo.6867552. This work is supported by the National Natural Science Foundation of China (41874197). S.Y.Li is also supported by the China Scholarship Council (award to S.Y. Li for 1 year study abroad at Ludwig Maximilians University Munich). EK is supported by German Research Foundation (DFG) under number KR 4375/2-1 within SPP "Dynamic Earth".

References

Antonova, E. E. (2004). Magnetostatic Equilibrium and Current Systems in the Earth's Magnetosphere. Streamers, Slow Solar Wind, and the Dynamics of the Magnetosphere, 33(5), 752-760. http://doi.org/10.1016/S0273-1177(03)00636-7

Antonova, E. E., Kirpichev, I. P., & Stepanova, M. V. (2014). Plasma Pressure Distribution in the Surrounding the Earth Plasma Ring and Its Role in the Magnetospheric Dynamics. Journal of Atmospheric and Solar-Terrestrial Physics, 115, 32-40. http://doi.org/10.1016/j.jastp.2013.12.005

Brandt, P. C. s., Roelof, E. C., Ohtani, S., Mitchell, D. G., & Anderson, B. (2004). Image/Hena: Pressure and Current Distributions During the 1 October 2002 Storm. Advances in Space Research, 33(5), 719-722. http://doi.org/10.1016/s0273-1177(03)00633-1

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). Classification And Regression Trees (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Camporeale, E. (2019). The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. Space Weather-the International Journal of Research and Applications, 17(8), 1166-1207. http://doi.org/10.1029/2018sw002061

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine learning, 20(3), 273-297. http://doi.org/Doi 10.1007/Bf00994018

Daly, P., & Kronberg, E. (2018). User Guide to the Rapid Measurements in the Cluster Science Archive (CSA): Version 5.2, Tech. Rep. CAA-EST-UG-RAP, European Space Agency, Paris.

De Michelis, P., Daglis, I. A., & Consolini, G. (2013). Average Terrestrial Ring Current Derived from Ampte/Cce-Chem Measurements. Journal of Geophysical Research: Space Physics, 102(A7), 14103-14111. http://doi.org/10.1029/96ja03743

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. The Annals of statistics, 32(2), 407-499.

Escoubet, C., Schmidt, R., & Goldstein, M. (1997). Cluster-Science and Mission Overview. The cluster and phoenix missions, 11-32.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. Journal of computer and system sciences, 55(1), 119-139. http://doi.org/DOI 10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of statistics, 29(5), 1189-1232. http://doi.org/DOI 10.1214/aos/1013203451

Gabrielse, C., V. Angelopoulos, A. Runov, and D. L. Turner (2014), Statistical characteristics of particle injections throughout the equatorial magnetotail, J. Geophys. Res. Space Physics, 119,2512–2535, doi:10.1002/2013JA019638.

Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems: " O'Reilly Media, Inc.".

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely Randomized Trees. Machine learning, 63(1), 3-42. http://doi.org/10.1007/s10994-006-6226-1

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression - Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55-&. http://doi.org/Doi 10.1080/00401706.1970.10488634

Ho, T. K., "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

Ilie, R., M. W. Liemohn, G. Toth, and R. M. Skoug (2012), Kinetic model of the inner magnetosphere with arbitrary magnetic field, J. Geophys. Res., 117, A04208, doi:10.1029/2011JA017189.

Iyemori, T., Takeda, M., Nose, M., Odagi, Y., & Toh, H. (2010). Mid-Latitude Geomagnetic Indices "Asy" and "Sym" for 2009 (Provisional). Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University, Japan.

Ke, G. L., Meng, Q., Finley, T., Wang, T. F., Chen, W., Ma, W. D., et al. (2017). Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (Nips 2017), 30. Retrieved from <Go to ISI>://WOS:000452649403021

Keika, K., Nose, M., Ohtani, S., Takahashi, K., Christon, S. P., & McEntire, R. W. (2005). Outflow of Energetic Ions from the Magnetosphere and Its Contribution to the Decay of the Storm Time Ring Current. Journal of Geophysical Research-Space Physics, 110, A09210, doi:10.1029/2004JA010970.

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.

Kistler, L. M., Mobius, E., Baumjohann, W., Paschmann, G., & Hamilton, D. C. (1992). Pressure Changes in the Plasma Sheet During Substorm Injections. Journal of Geophysical Research-Space Physics, 97(A3), 2973-2983. http://doi.org/Doi 10.1029/91ja02802

Kistler, L. M., & Mouikis, C. G. (2016). The Inner Magnetosphere Ion Composition and Local Time Distribution over a Solar Cycle. Journal of Geophysical Research: Space Physics, 121(3), 2009-2032. http://doi.org/10.1002/2015ja021883

Kronberg, E. A., & Daly, P. W. (2013). Spectral Analysis for Wide Energy Channels. Geoscientific Instrumentation, Methods and Data Systems, 2(2), 257-261.

Kronberg, E. A., Daly, P. W., Dandouras, I., Haaland, S., & Georgescu, E. (2010). Generation and Validation of Ion Energy Spectra Based on Cluster Rapid and Cis Measurements. In The Cluster Active Archive (pp. 301-306): Springer.

Kronberg, E. A., Daly, P. W., & Vilenius, E. (2022). Calibration Report of the RAPID Measurements in the Cluster Science Archive (CSA), technical report. European Space Agency.

Kronberg, E. A., Gastaldello, F., Haaland, S., Smirnov, A., Berrendorf, M., Ghizzardi, S., et al. (2020). Prediction and Understanding of Soft-Proton Contamination in Xmm-Newton: A Machine Learning Approach. The Astrophysical Journal, 903(2), 89. http://doi.org/10.3847/1538-4357/abbb8f.

Kronberg, E. A., Grigorenko, E., Haaland, S., Daly, P. W., Delcourt, D. C., Luo, H., et al. (2015). Distribution of Energetic Oxygen and Hydrogen in the near‐Earth Plasma Sheet. Journal of Geophysical Research: Space Physics, 120(5), 3415-3431.

Kronberg, E. A., Hannan, T., Huthmacher, J., Münzer, M., Peste, F., Zhou, Z., et al. (2021). Prediction of Soft Proton Intensities in the near-Earth Space Using Machine Learning. The Astrophysical Journal, 921(1), 76.

Kronberg, E. A., Welling, D., Kistler, L. M., Mouikis, C., Daly, P. W., Grigorenko, E. E., et al. (2017). Contribution of Energetic and Heavy Ions to the Plasma Pressure: The 27 September to 3 October 2002 Storm. Journal of Geophysical Research: Space Physics, 122(9), 9427-9439. http://doi.org/10.1002/2017ja024215

Laakso, H., Perry, C., McCaffrey, S., Herment, D., Allen, A., Harvey, C., et al. (2010). Cluster Active Archive: Overview. The cluster active archive, 3-37.

Lui, A. T. Y. (2003). Inner Magnetospheric Plasma Pressure Distribution and Its Local Time Asymmetry. Geophysical Research Letters, 30(16).

Lui, A. T. Y., & Hamilton, D. C. (1992). Radial Profiles of Quiet Time Magnetospheric Parameters. Journal of Geophysical Research, 97(A12), 19325. http://doi.org/10.1029/92ja01539

Luo, H., Kronberg, E. A., Nykyri, K., Trattner, K. J., Daly, P. W., Chen, G. X., et al. (2017). Imf Dependence of Energetic Oxygen and Hydrogen Ion Distributions in the near‐Earth Magnetosphere. Journal of Geophysical Research: Space Physics, 122(5), 5168-5180.

http://doi.org/10.1002/2016ja023471
Meng, C. I., Lui, A., Krimigis, S., Ismail, S., & Williams, D. (1981). Spatial Distribution of Energetic Particles in the Distant Magnetotail. Journal of Geophysical Research: Space Physics, 86(A7), 5682-5700.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python. the Journal of machine Learning research, 12, 2825-2830.
Reme, H., Aoustin, C., Bosqued, J., Dandouras, I., Lavraud, B., Sauvaud, J., et al. (2001). First Multispacecraft Ion Measurements in and near the Earth's Magnetosphere with the Identical Cluster Ion Spectrometry (CIS) Experiment. Ann. Geophys., 19, 1303–1354, https://doi.org/10.5194/angeo-19-1303-2001, 2001.
Roederer, J. G. (1967). On the Adiabatic Motion of Energetic Particles in a Model Magnetosphere. Journal of Geophysical Research, 72(3), 981-992.
Roederer, J. G., & Lejosne, S. (2018). Coordinates for Representing Radiation Belt Particle Flux. Journal of Geophysical Research: Space Physics, 123(2), 1381-1387. http://doi.org/10.1002/2017ja025053
Sarafopoulos, D., Sidiropoulos, N., Sarris, E., Lutsenko, V., & Kudela, K. (2001). The Dawn-Dusk Plasma Sheet Asymmetry of Energetic Particles: An Interball Perspective. Journal of Geophysical Research: Space Physics, 106(A7), 13053-13065.
Sergeev, V. A., Pulkkinen, T. I., Pellinen, R. J., & Tsyganenko, N. A. (1994). Hybrid State of the Tail Magnetic Configuration During Steady Convection Events. Journal of Geophysical Research, 99(A12), 23571. http://doi.org/10.1029/94ja01980
Smirnov, A., Berrendorf, M., Shprits, Y., Kronberg, E. A., Allison, H. J., Aseev, N. A., et al. (2020). Medium Energy Electron Flux in Earth's Outer Radiation Belt (Merlin): A Machine Learning Model. Space Weather, 18(11), e2020SW002532.
Smirnov, A. G., Kronberg, E. A., Daly, P. W., Aseev, N. A., Shprits, Y. Y., & Kellerman, A. C. (2020). Adiabatic Invariants Calculations for Cluster Mission: A Long-Term Product for Radiation Belts Studies. Journal of Geophysical Research: Space Physics, 125(2), e2019JA027576.
Stepanova, M., Antonova, E. E., & Bosqued, J. M. (2008). Radial Distribution of the Inner Magnetosphere Plasma Pressure Using Low-Altitude Satellite Data During Geomagnetic Storm: The March 1–8, 1982 Event. Advances in Space Research, 41(10), 1658-1665. http://doi.org/10.1016/j.asr.2007.06.002
Stepanova, M., Antonova, E. E., Moya, P. S., Pinto, V. A., & Valdivia, J. A. (2019). Multisatellite Analysis of Plasma Pressure in the Inner Magnetosphere During the 1 June 2013 Geomagnetic Storm. Journal of Geophysical Research: Space Physics, 124(2), 1187-1202. http://doi.org/10.1029/2018ja025965
Stephens, G. K., Sitnov, M. I., Kissinger, J., Tsyganenko, N. A., McPherron, R. L., Korth, H., & Anderson, B. J. (2013). Empirical Reconstruction of Storm Time Steady Magnetospheric Convection Events. Journal of Geophysical Research: Space Physics, 118(10), 6434-6456. http://doi.org/10.1002/jgra.50592
Tapping, K. (2013). The 10.7 Cm Solar Radio Flux (F10. 7). Space Weather, 11(7), 394-406.
Tsyganenko, N. A. (1989). A Magnetospheric Magnetic Field Model with a Warped Tail Current Sheet. Planetary and Space Science, 37(1), 5-20.
Wing, S., & Newell, P. T. (1998). Central Plasma Sheet Ion Properties as Inferred from Ionospheric Observations. Journal of Geophysical Research: Space Physics, 103(A4), 6785-6800. http://doi.org/10.1029/97ja02994
Wilken, B., Daly, P., Mall, U., Aarsnes, K., Baker, D., Belian, R., et al. (2001). First Results from the Rapid Imaging Energetic Particle Spectrometer on Board Cluster. Ann. Geophys., 19, 1355–1366, https://doi.org/10.5194/angeo-19-1355-2001, 2001.

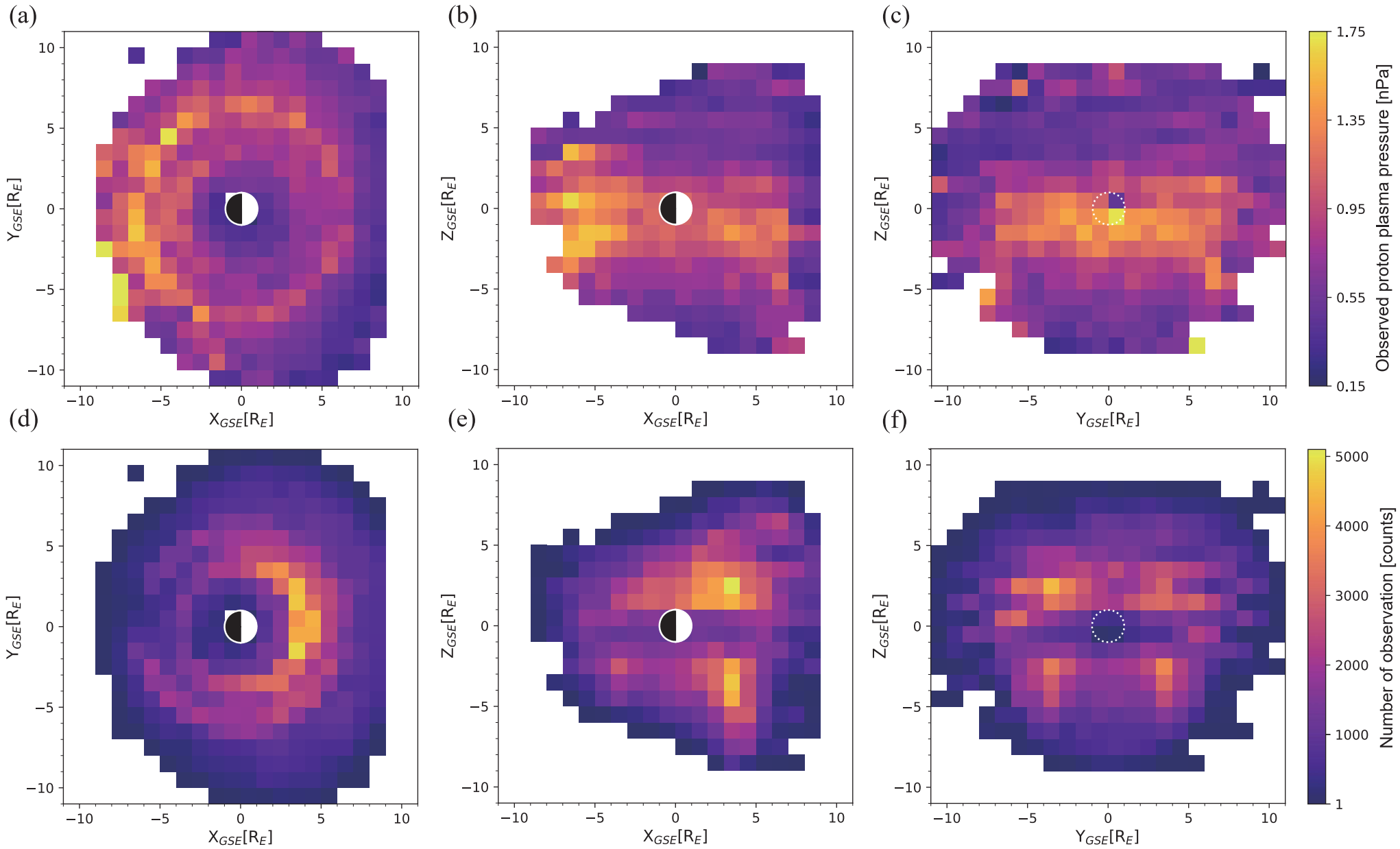

Figure 1. (a-c) Distributions of the observed H^+ plasma pressures by SC4 from February 2001 to February 2018 in the GSE coordinate system. (d-f) Distributions of the number of measurements corresponding to (a-c). Resolution (bin size) is 1 Re.

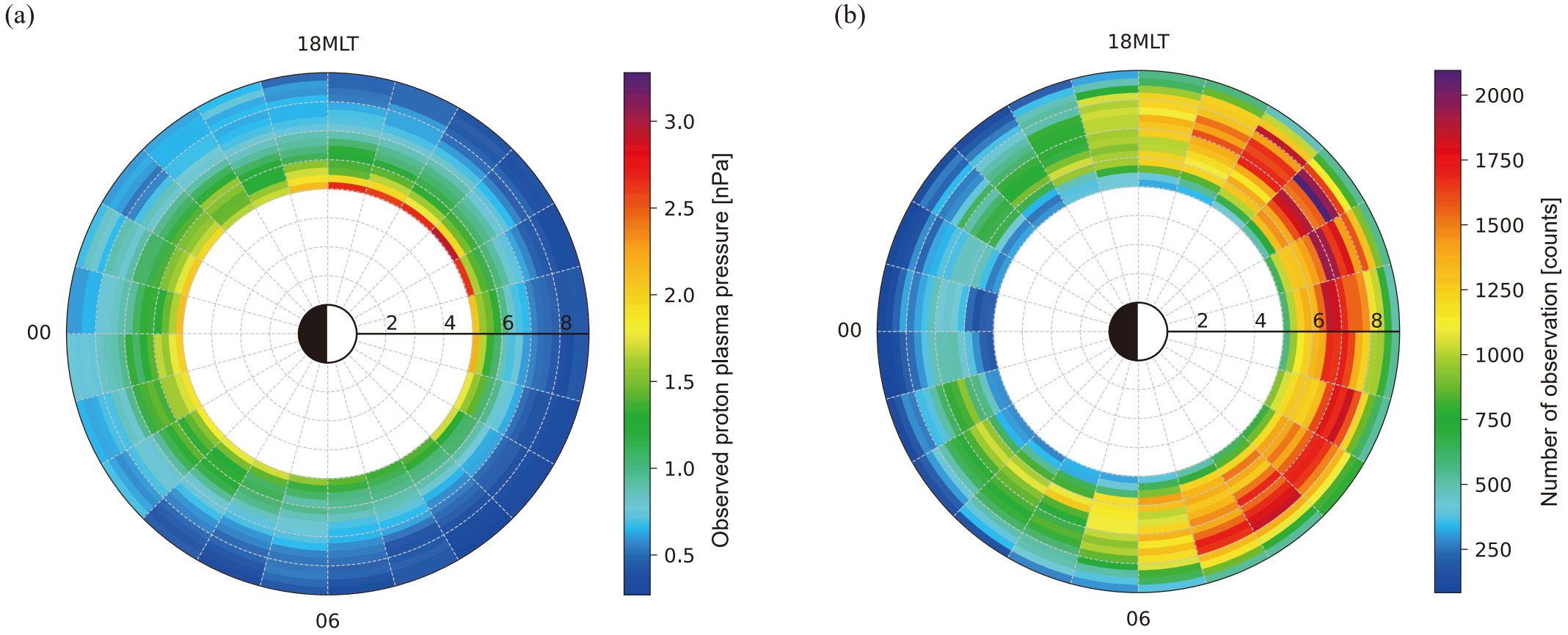

Figure2. Distributions of the observed H^+ plasma pressures (left) and the number of measurements (right) by SC4 from February 2001 to February 2018 in the L*-MLT coordinate. The L* resolution (bin size) is 0.25. MLT resolution (bin size) is 1.

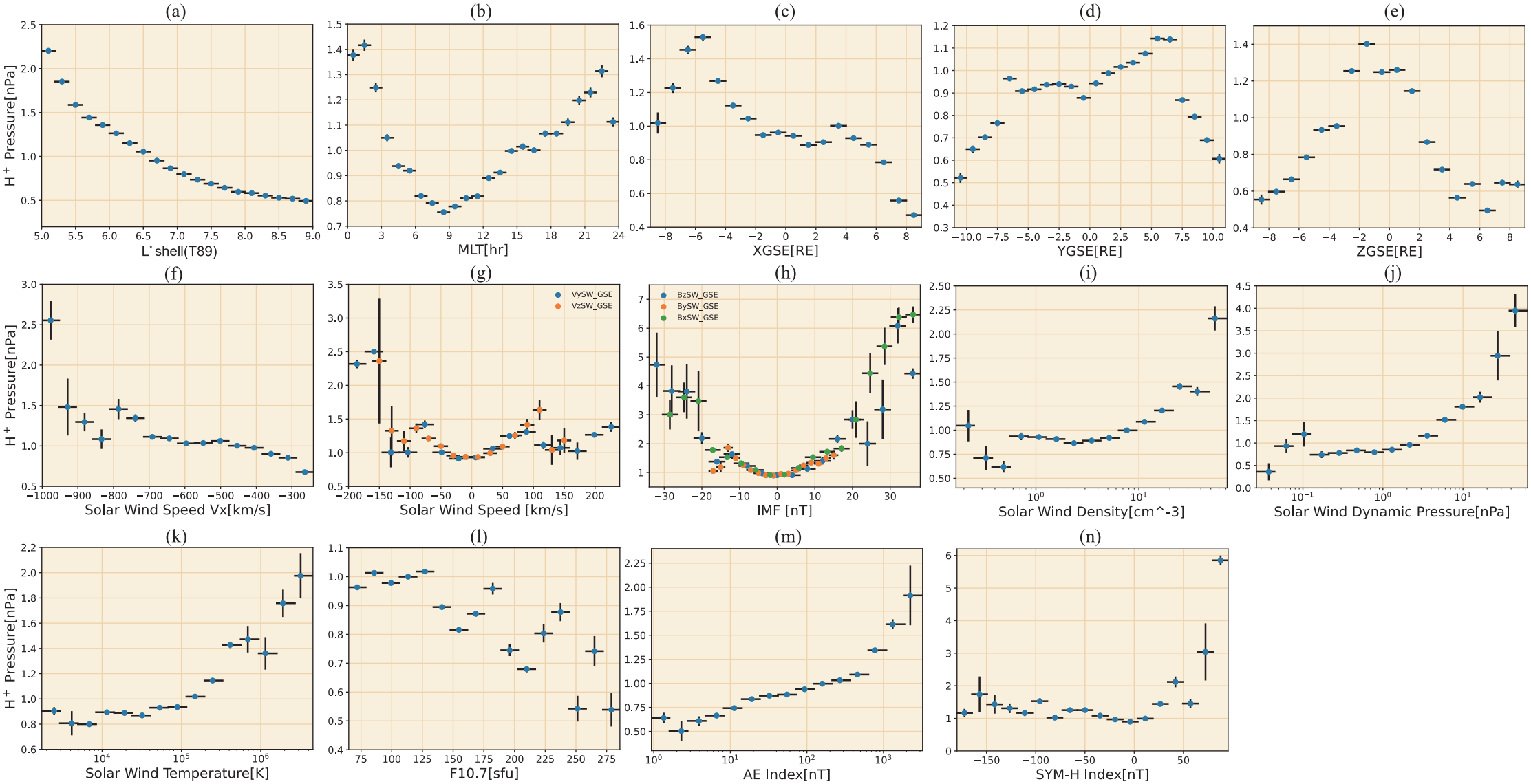

Figure3. Relations of mean H^+ plasma pressure and (a)-(e) L^* shell, MLT, XGSE, YGSE, and ZGSE, respectively; (f)-(g) the solar wind Vx ,Vy and Vz components; (h) IMF components in GSE;(i)-(k) solar wind density, dynamic pressure and temperature, respectively; (l) F10.7 parameter; (m) AE index and (n) SYM-H index. Vertical lines represent confidence intervals at 95% confidence level.

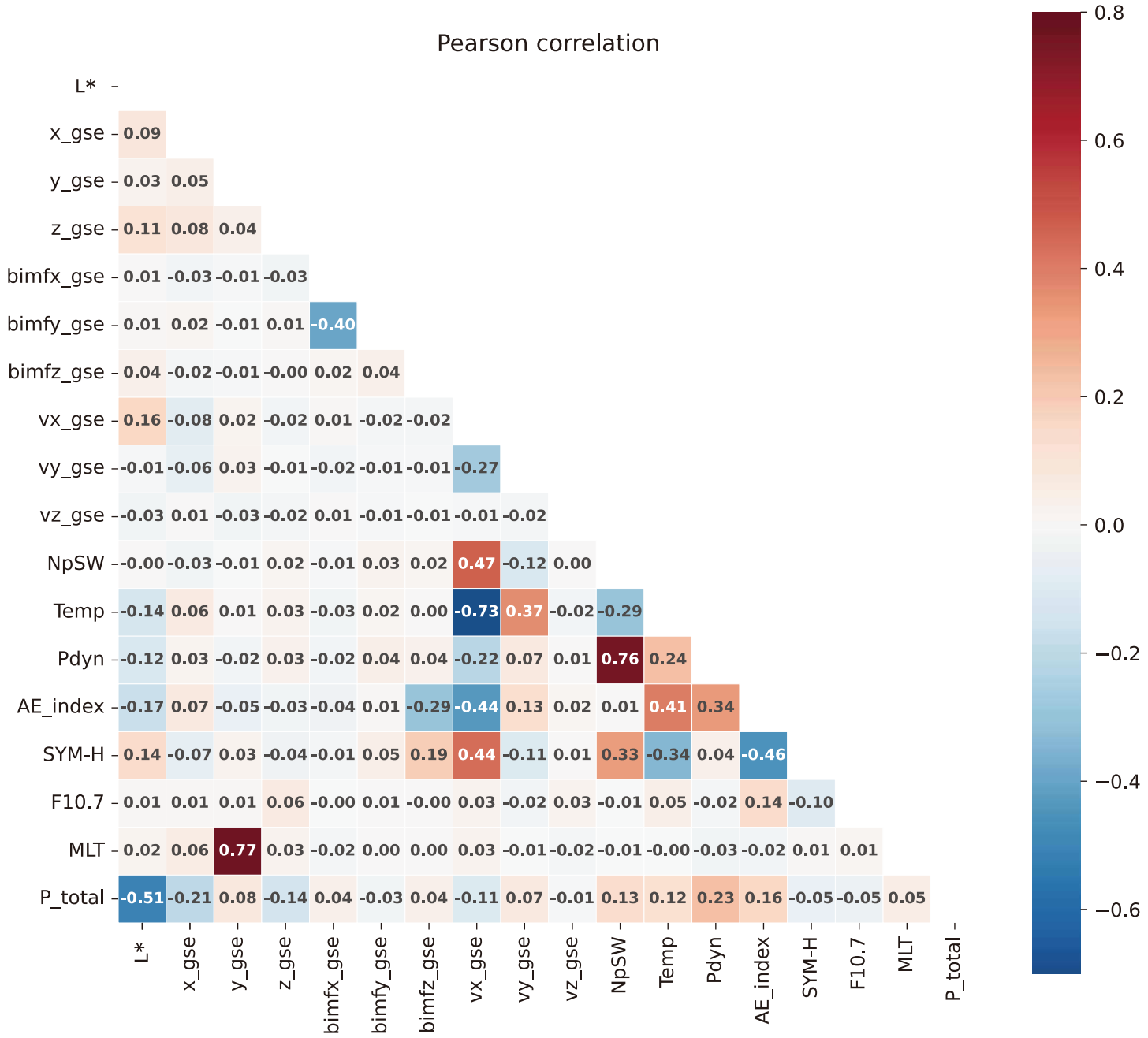

Figure4. Pearson correlation matrix between input parameters and H^+ plasma pressure.

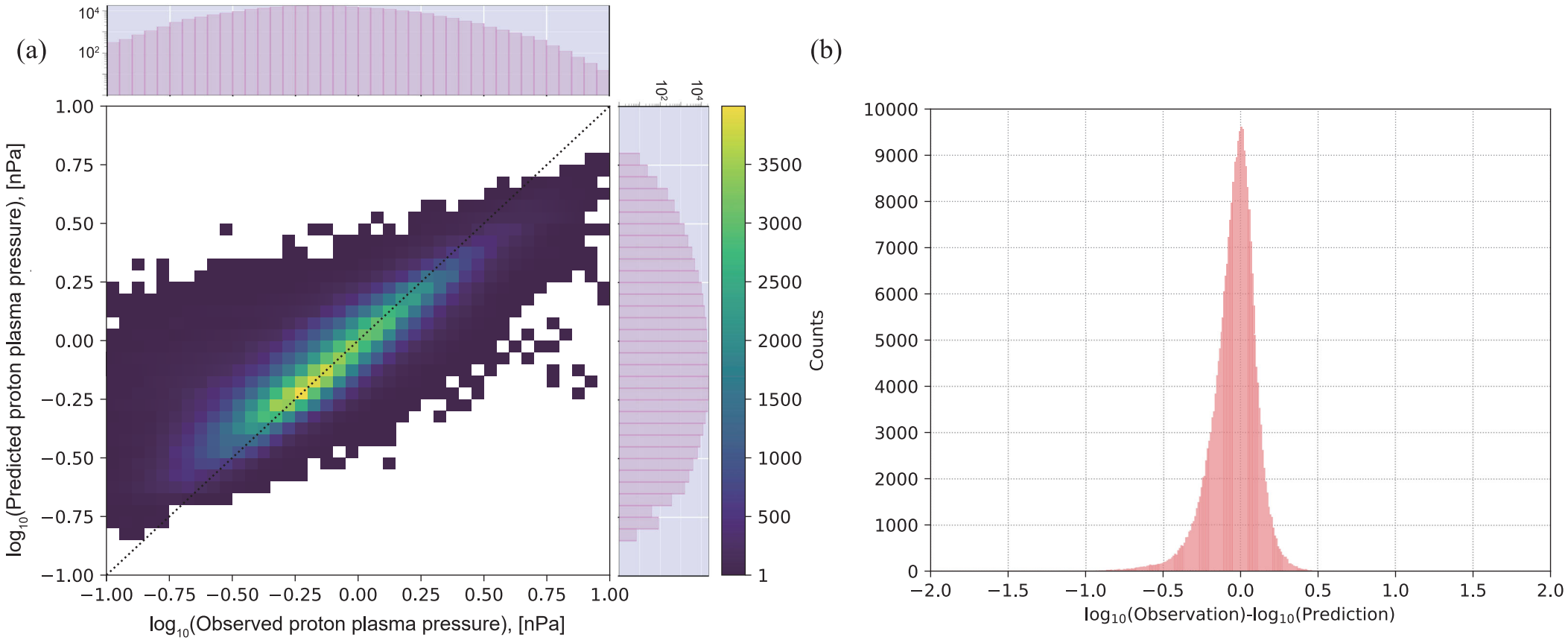

Figure 5. (a) The observed (x-axis) vs. the predicted (y-axis) proton plasma pressure from the training set. The color represents the number of samples in the corresponding bin (10^0.05). The diagonal shows the one-to-one ratio between the observed and predicted pressure.(b)The histogram of the model residuals.

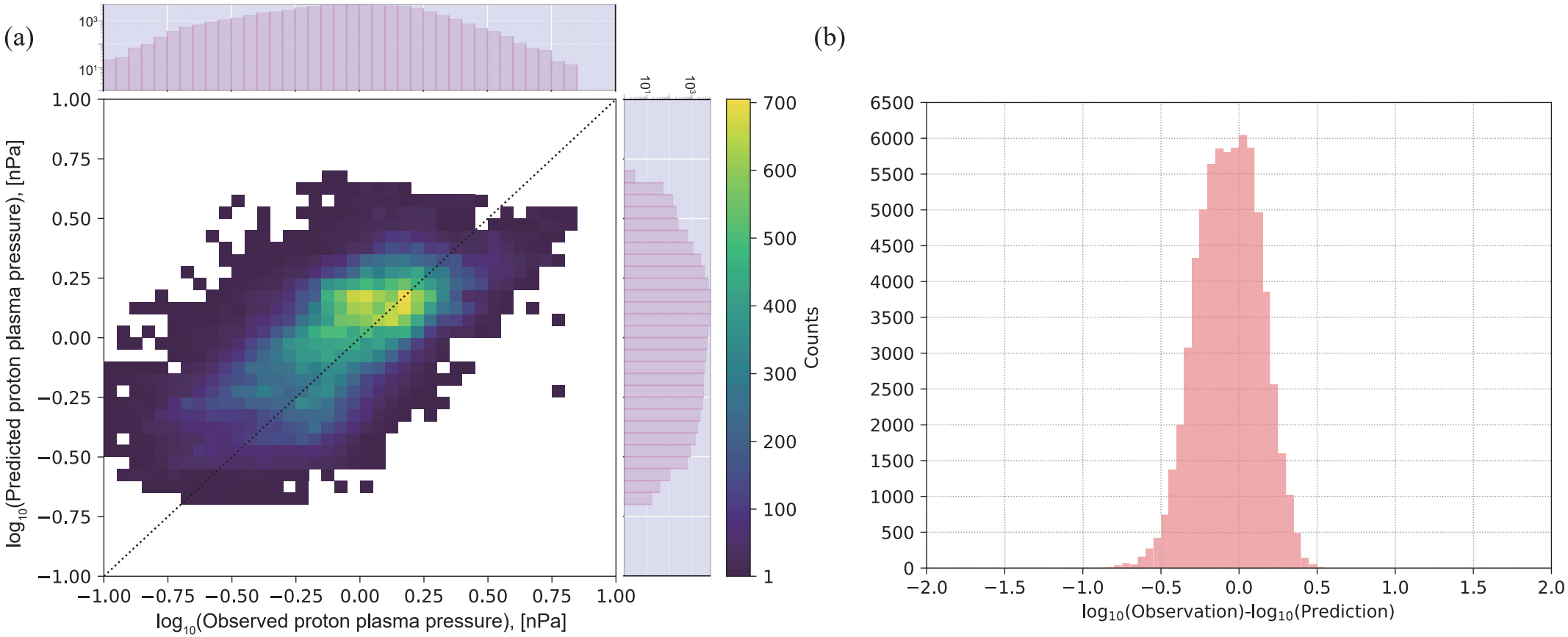

Figure 6. Results from the test set. Same as **Figure5**.

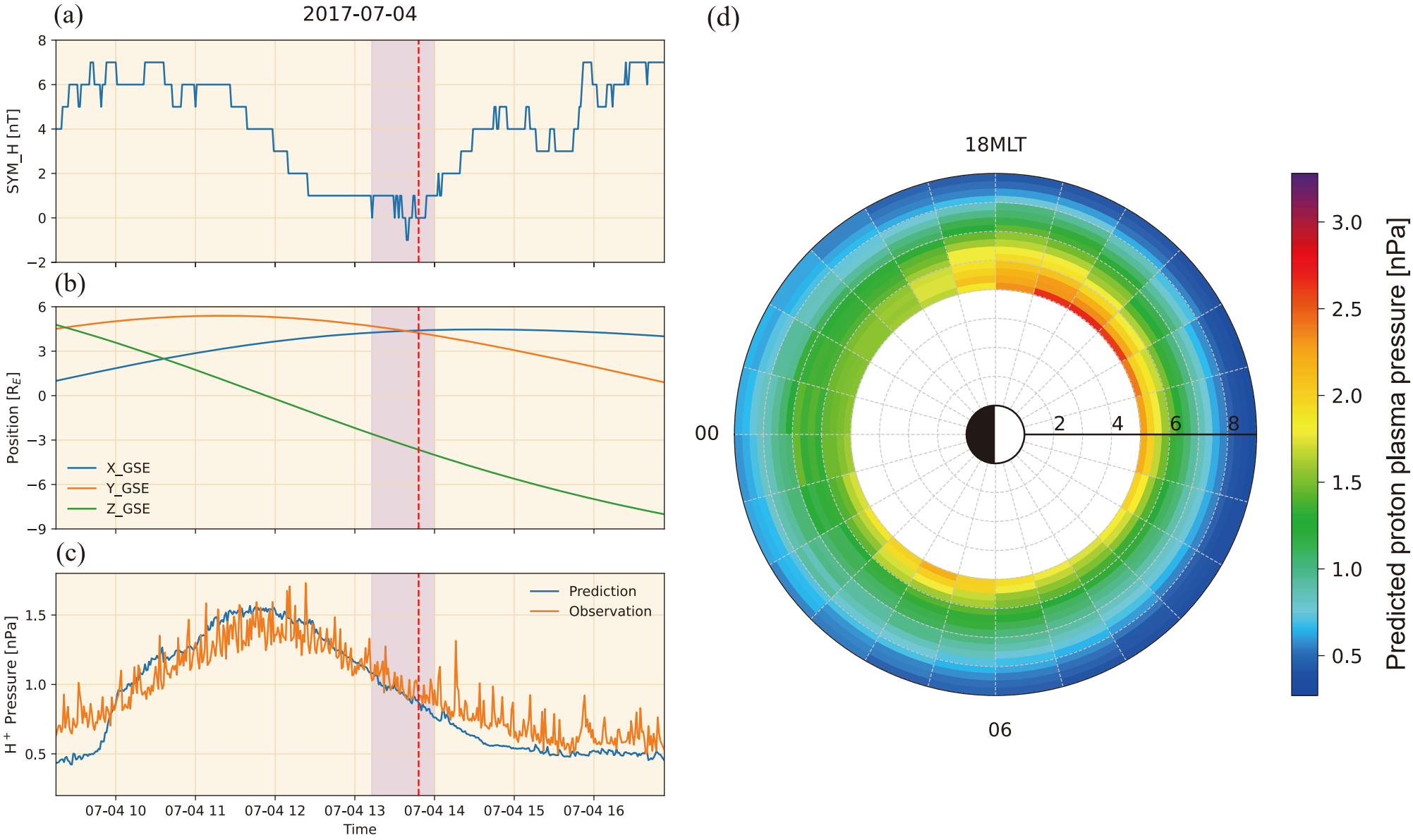

Figure 7. (a)SYM-H index, (b)position, (c)predicted pressure (blue) vs. measured pressure (orange) within the time interval on 2017-07-04 under quiet geomagnetic conditions. (d) Distribution of the predicted H^+ plasma pressure using median values of the parameters over the time period in the purple region as input predictors. MLT is given by the time indicated by the red dotted line.

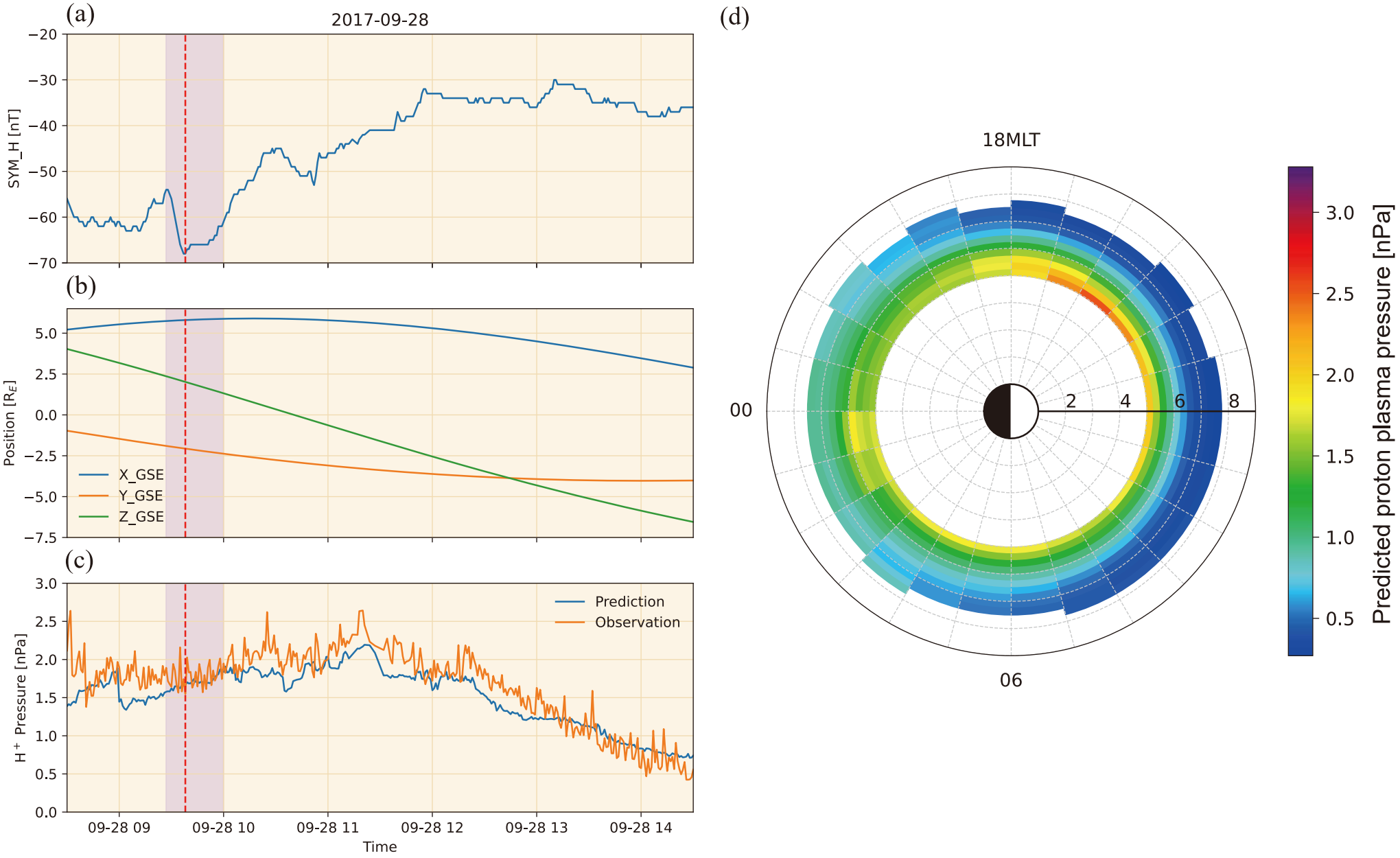

Figure 8. Results within the time interval on 2017-09-28 during the main phase of a geomagnetic storm. Others are the same as **Figure 7**.

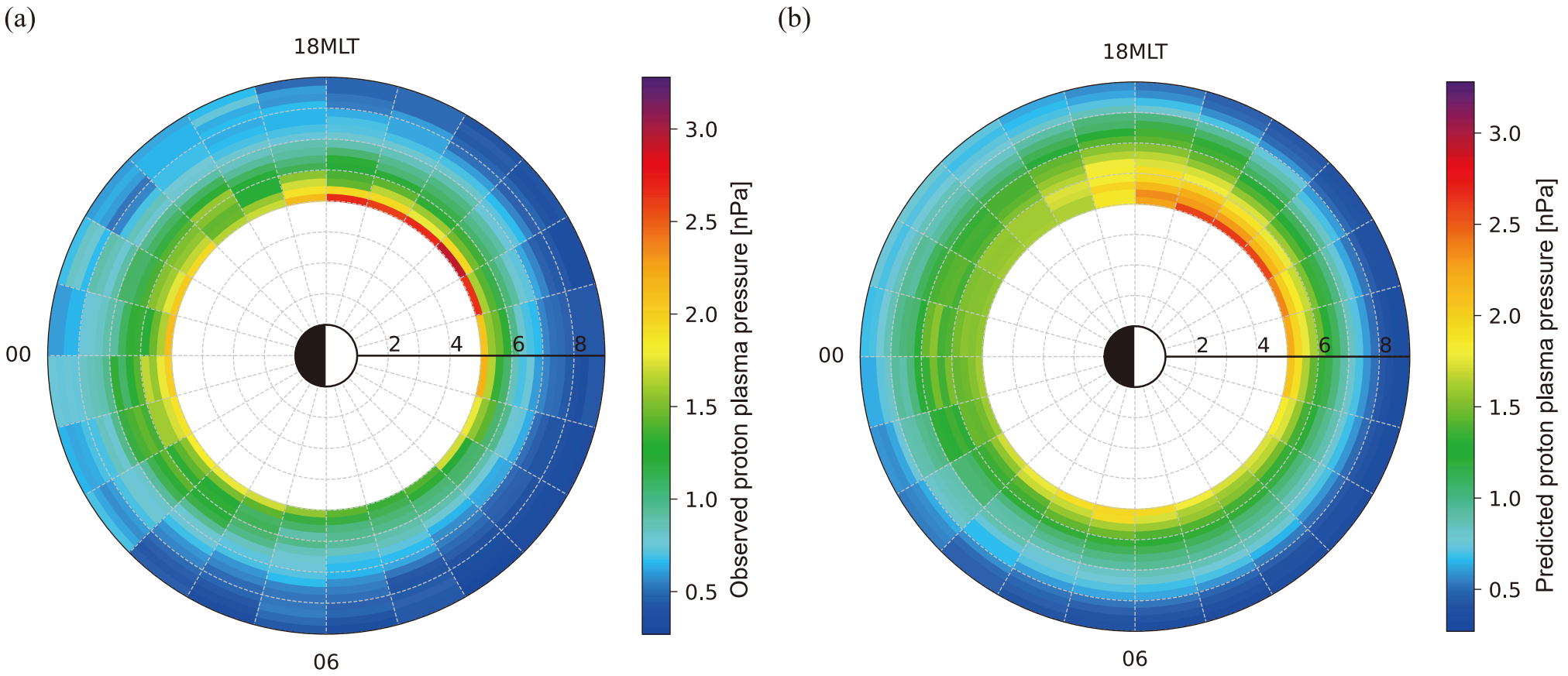

Figure9. (a) Distribution of the observed H^+ plasma pressures by SC4 from February 2001 to February 2018 in the L*-MLT coordinate (Same as **Figure 2a**). (b) Distribution of the predicted H^+ plasma pressures using median values of the parameters over the time range of February 2001 to February 2018 as input predictors in the L*-MLT coordinate.The L* resolution (bin size) is 0.25. MLT resolution (bin size) is 1.

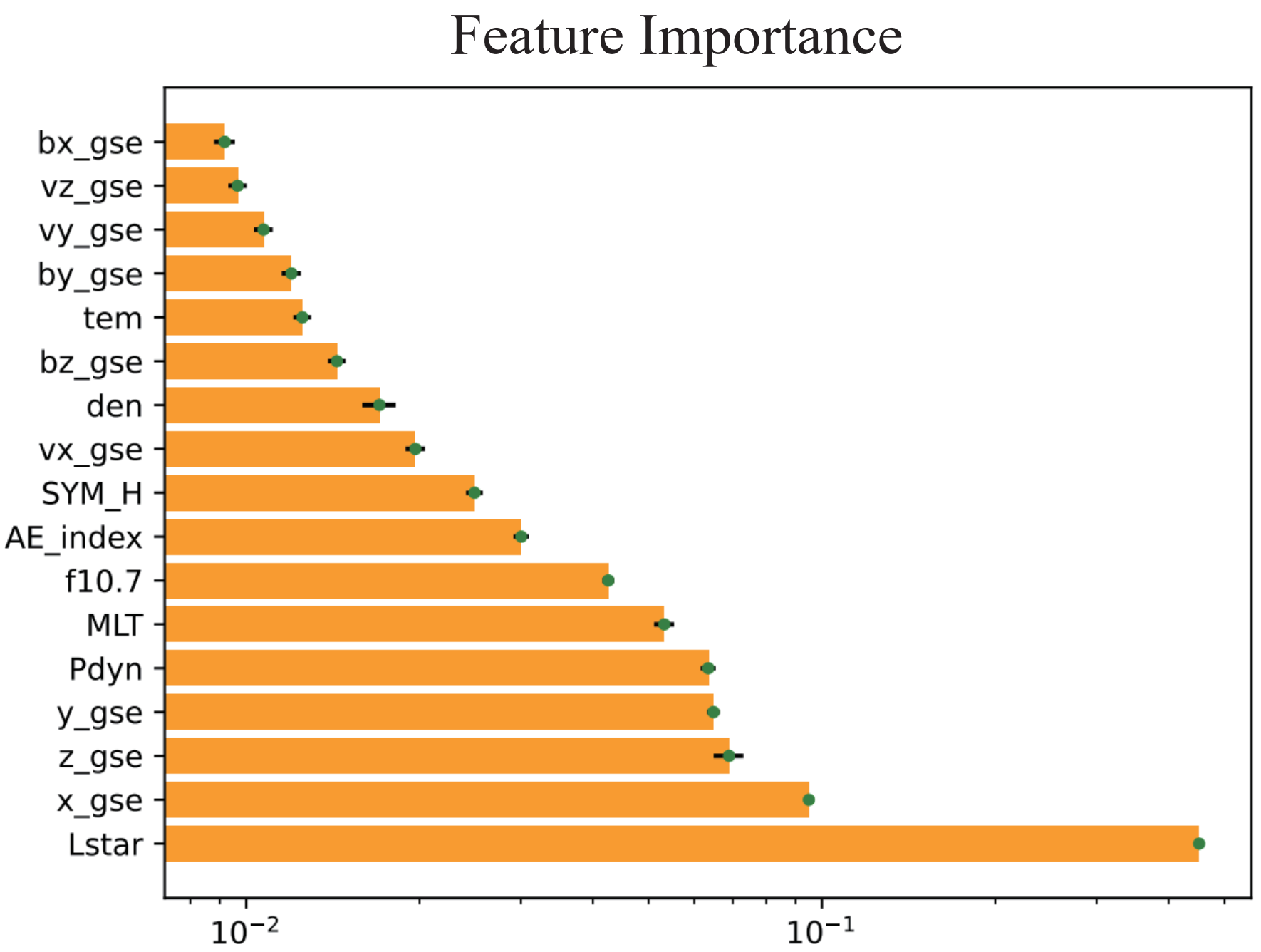

Figure 10. Importance of the different parameters in prediction of H^+ plasma pressure based on training data set. The black horizontal lines represent confidence intervals at 95% confidence level.

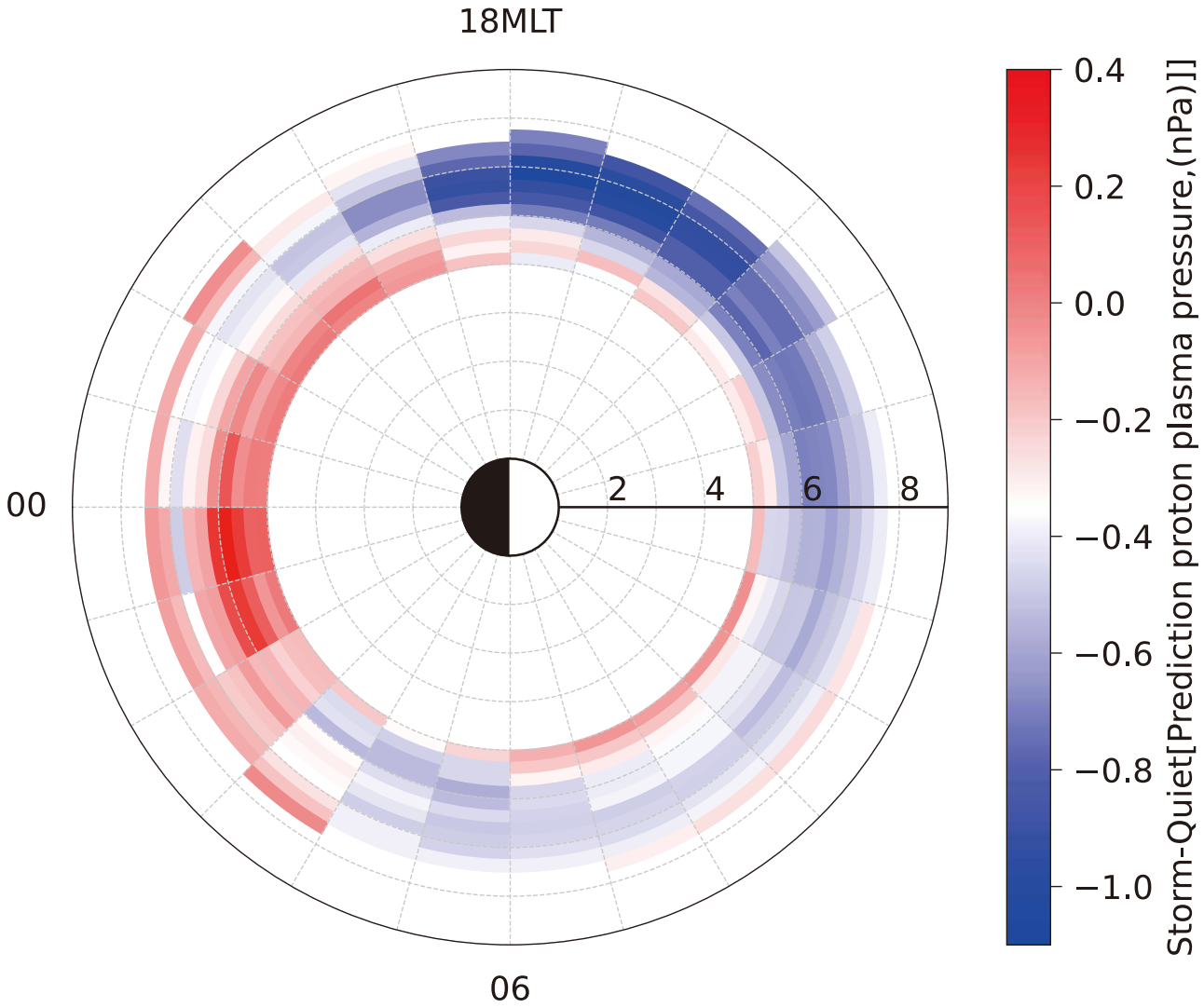

Figure 11. Distribution of the difference between the predicted H^+ plasma pressure under disturbed (shown in **Figure 8d**) and quiet (shown in **Figure 7d**) geomagnetic conditions.

Table 1. Overview of the input features

Feature	Unit	Description
L*		calculated with T89 model
MLT	hr	magnetic local time
x_gse, y_gse, z_gse	RE	position of Cluster in GSE coordinate
Vx/y/z_gse	km/s	components of the solar wind speed in GSE
Bimfx/y/z_gse	nT	components of the interplanetary magnetic field (IMF) in GSE
NpSW	/cm3	solar wind density
Temp	K	solar wind temperature
Pdyn	nPa	solar wind dynamic pressure
AE_index	nT	auroral electrojet index
SYM_H	nT	symmetric H-component index
F10.7	sfu	the solar radio flux at 10.7 cm

Table2. Size and periods of the data subsets after splitting. The time is given in UT.

Subset	Start	End	Number of Points
Train/Validation	2001-02-04 12:31:00	2016-05-16 17:08:00	269184
Test	2016-05-16 17:09:00	2018-02-18 00:02:00	67297

Table 3. Performance of Different Models with Default Input Values

Regressor	Train/Validation Spearman	Test Spearman
ExtraTrees	1.000	0.670
DecisionTree	1.000	0.469
RandomForest	0.996	0.618
LGBM	0.849	0.661
HistGradientBoosting	0.849	0.658
GradientBoosting	0.780	0.664
LinearSVR	0.652	0.634
RidgeRegression	0.648	0.642
LARSRegression	0.645	0.633
AdaBoost	0.628	0.611

Table 4. ExtraTrees Hyperparameters Used for Model Setup

Name	Search range	Value
n_estimators	10-350	140
max_depth	1-30	13
min_samples_leaf	1-50	35

Note. Other parameters use default values.

Table 5. Performance of the Model for Trained/Validation and Test Data Sets

Metrics	Train	Validation	Test
Spearman	0.89	0.71	0.68
MSE	0.12	0.39	0.31
MAE	0.21	0.39	0.40
r^2	0.79	0.13	0.26

Appendix

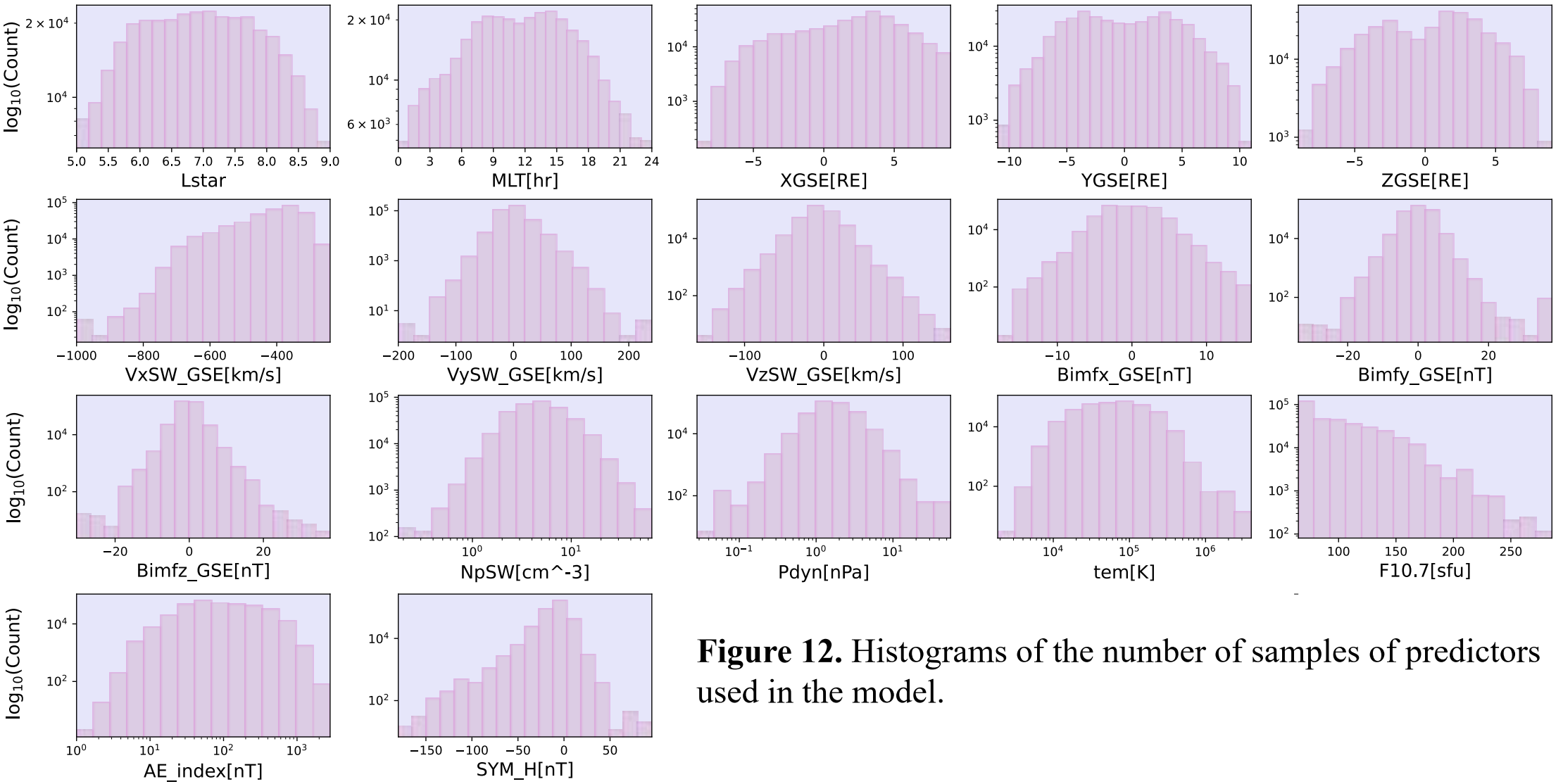

Figure 12. Histograms of the number of samples of predictors used in the model.

Table 6. Comparison of median values of predictors during different time periods

Parameter	Median of quiet time	Median of storm time	Median of all points
SYM_H index (nT)	1	-65.5	-8
Bx_gse (nT)	0.215	-2.49	-0.23
By_gse (nT)	-2.675	0.83	0.11
Bz_gse (nT)	-0.54	3.41	-0.01
Vx_gse (km/s)	-342.8	-687.05	-403.2
Vy_gse (km/s)	-6.6	28.5	-1.2
Vz_gse (km/s)	4.75	-28.9	-5.5
Density (/cm^3)	7.81	3.84	4.7
Temperature (K)	15381.5	263686	63461
Dynamic pressure (nPa)	1.835	3.675	1.64
AE index (nT)	118	581.5	87
F10.7 (sfu)	74.2	91.2	94.6

Prediction of Plasma Pressure in the Outer Part of the Inner Magnetosphere using Machine Learning

S. Y. Li[1,2,3,4], E. A. Kronberg[4,*], C. G. Mouikis[5], H. Luo[1,2,3,], Y. S. Ge[1,2,3,], A. M. Du[1,2,3,]

[1]CAS Engineering Laboratory for Deep Resources Equipment and Technology, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China

[2]Innovation Academy for Earth Science, CAS, Beijing 100029, China

[3]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[4]Department of Earth and Environmental Sciences (Geophysics), Ludwig Maximilian University of Munich (LMU) Munich, Theresienstr. 41, Munich, 80333, Germany

[5]Department of Physics and Space Science Center, University of New Hampshire, Durham, New Hampshire, USA

*Corresponding Author: kronberg@geophysik.uni-muenchen.de

Key Points:

1. A machine learning model is created to predict 3-D distribution of proton plasma pressures at L*=5-9 for energies ~40eV-4MeV
2. Our model based on Extra-Trees Regressor reproduces well the global distributions as well as the pressure along a spacecraft trajectory
3. The results of our model are helpful for the interpretation of the plasma pressure in the outer part of the magnetosphere

Plain Language Summary

The distribution of the plasma pressures in the magnetosphere is a key parameter for the assessment of the magnetostatic equilibrium, the dynamics of geomagnetic storms, and the magnetospheric electric current system. In addition, the outer part of the inner magnetosphere (L^*=5-9) is often used as the boundary in the inner magnetosphere simulations, where the initial composition is specified. Thus, the distribution of the plasma pressure at L^*=5-9 is essential for the simulations of the inner magnetosphere and understanding of the underlying magnetospheric dynamic processes. Although, there are many previous studies on the distribution of plasma pressures, building a model to predict the 3-D distribution of plasma pressures remains challenging. Based on 17 years of data from both CIS and RAPID instruments onboard the Cluster spacecraft mission, a machine-learning-based model for predicting proton pressures at energies from ~40eV to 4MeV in the outer part of the inner magnetosphere (L^*=5-9) is built. We set up the 3-D model for the prediction of the proton pressures depending on the location, solar, solar wind, and geomagnetic activity indices. The model gives reliable predictions and can be used for the interpretation of the dynamics of the inner magnetosphere under different geomagnetic conditions which can also provide valuable guidance to the space weather (such as magnetic storms) forecast.

Abstract

The information on plasma pressures in the outer part of the inner magnetosphere is important for simulations of the inner magnetosphere and the better understanding of its dynamics. Based on 17-year observations from both CIS and RAPID instruments onboard the Cluster mission, we used machine-learning-based models to predict proton plasma pressures at energies from ~40eV to 4MeV in the outer part of the inner magnetosphere (L^*=5-9). The location in the magnetosphere, and parameters of solar, solar wind, and geomagnetic activity from the OMNI database are used as predictors. We trained several different machine-learning-based models and compared their performances with observations. The results demonstrate that the Extra-Trees Regressor has the best predicting performance. The Spearman correlation between the observations and predictions by the model data is about 68%. The most important parameter for predicting proton pressures in our model is the L^* value, which is related to the location and distance. The most important predictor of solar, solar wind, and geomagnetic activity is the

solar wind dynamic pressure. Based on the observations and predictions by our model, we find that no matter under quiet or disturbed geomagnetic conditions, both the dusk-dawn asymmetry at the dayside with higher pressures at the duskside and the day-night asymmetry with higher pressures at the nightside occur. Our results have direct practical applications, for instance, inputs for simulations of the inner magnetosphere or the reconstruction of the 3-D magnetospheric electric current system based on the magnetostatic equilibrium, and can also provide valuable guidance to the space weather forecast.

1. Introduction

In the inner magnetosphere, the plasma pressure plays a key role in the understanding of the main magnetospheric dynamic processes. The knowledge about distributions of the plasma pressures under different geomagnetic conditions is necessary for explaining how the Earth's magnetosphere reaches the magnetostatic equilibrium (the plasma pressure gradient is compensated by Ampere's force) and what specific conditions are necessary to maintain it (Antonova, 2004; Stepanova et al., 2019). In addition, one of the key parameters for understanding the evolution of geomagnetic storms and substorms is the plasma pressure distribution in the inner magnetosphere (Kronberg et al., 2017; Stepanova et al., 2008). The increase of the inner magnetospheric plasma pressure is one of the main features of magnetic storms (Stepanova et al., 2019).

The distribution of plasma pressures in the inner magnetosphere has been studied extensively during last decades. Based on the measurements from the high-altitude AMPTE/CCE satellite, Lui and Hamilton (1992) obtained average radial profiles of plasma pressures from a case study during geomagnetically quiet conditions. These profiles showed a peak generally at L = 3 to 4 and decreased from L = 4 to L = 9 rather monotonically. Using data from the same satellite, De Michelis et al. (2013) presented the statistical study of plasma pressure profiles which were averaged over more than 2 years data. The low-activity pressure profile gave the same general features as the profiles in the case study published by Lui and Hamilton (1992). The disturbed pressure profile had a peak at a higher L value (L~4.5) and decreased from L = 5 to L = 8, also rather monotonically. The equatorial plasma pressure distribution can be obtained from low-altitude measurements under the assumption that the plasma pressure is conserved along a magnetic field line (Wing & Newell, 1998). Based on the energetic neutral

atom (ENA) images obtained by HENA onboard the IMAGE spacecraft, Brandt et al. (2004) inferred the evolution of the global plasma pressure distribution during storms, showing that there was a peak of the proton pressure located around the midnight. Similarly, Lui (2003) found that the proton pressures were generally higher in the dusk-midnight sector than in the post-midnight sector under disturbed geomagnetic conditions within the L shells from 2 to 9 in the equatorial region. Using the data from both low-orbiting (DMSP 16–18 spacecraft, satellites NOAA 15–19, and METOP 1–2 satellites) and high-orbiting satellites (the THEMIS and Van Allen Probes), Stepanova et al. (2019) performed a multisatellite analysis of the variation of plasma pressures near the equatorial plane between 7 to 13 R_E during a strong geomagnetic storm. They also found that the plasma pressure inside the magnetosphere is mainly controlled by the solar wind dynamic pressure.

Among these previous studies, most of them focus on the 2-D equatorial plasma pressure distribution (Antonova et al., 2014; Lui, 2003; Stepanova et al., 2019; Wing & Newell, 1998). The 3-D plasma pressure distribution in the inner magnetosphere is relatively unknown. This kind of 3-D distribution is not only helpful to understand the dynamics of the inner magnetosphere, but also important for the simulations of the inner magnetosphere. For example, it can be used to deduce the distribution of the temperature which is an important input for some simulations models (such as Hot Electron Ion Drift Integrator (HEIDI) model (Ilie et al., 2012)). In addition, using the steady−state force balance equation $\nabla P = j \times B$ (Sergeev et al., 1994; Stephens et al., 2013) allows one to reconstruct the 3-D electric current system with the 3-D plasma pressure distribution.

In this study, we derive a predictive model for the proton pressures at energies from ~40eV to 4MeV in the outer part of the inner magnetosphere (L^*=5-9). For this energy range, we combined the data from both CIS and RAPID instruments onboard Cluster. This L^* range is selected because it is the region that is often used as the boundary in the inner magnetosphere simulations, where the initial composition is specified (Kistler & Mouikis, 2016). Instrumentally, we restricted the minimum distance to L^*=5 to reduce the contamination of the results by energetic electrons from the outer radiation belt. To enable modeling of the complex non-linear multidimensional dependencies, we trained several different machine-learning-based models and compared their performances with the observations. Moreover, by

using a machine learning method we can utilize the full range of solar or geomagnetic parameters as inputs to infer and analyze the plasma pressure distribution instead of only considering a few of them as most previous studies did. This helps to increase the performance of the predictions.

To summarize, our study aims are to (1) test the capability of different machine learning algorithms and to present the best one for the prediction of the proton plasma pressures in the outer part of the inner magnetosphere; (2) reveal which parameters are the most important for the prediction of proton plasma pressures; (3) compare and analyze the prediction of the proton plasma pressure distributions under different geomagnetic conditions and (4) help future studies that require a proton plasma pressure model. The remainder of this paper is organized as follows. In section 2 we describe the observations and data analysis. Section 3 is concerned with the methodology used for this study. Section 4 presents the results. The discussions and conclusions are drawn in the last two sections 5 and 6.

2. Observations and Data Set

In this section, we first introduce the data and the method we used for calculating proton plasma pressures. Then, the predictors, also called features, (the variables that are potentially capable to predict the proton pressures) are also discussed.

2.1 Instrumentation and Data

The Cluster mission consists of four identical spacecrafts, each carrying 11 instruments. The satellites were launched in two pairs in late 2000, and after a 6-months commissioning phase, the mission moved into an operational phase in February 2001 (Laakso et al., 2010). For the first ~6 years of the mission, the spacecrafts were placed into a highly elliptical, polar orbits apogee at 19.6 Re and perigee at 4 Re (Escoubet et al., 1997). The orbit has evolved over time, passing through the inner magnetosphere closer to the equator. It also covers the full range of local times over the course of a year (Kistler & Mouikis, 2016). Thus, the data from the Cluster mission is proper for the study of 3-D distribution of proton pressures in the inner magnetosphere. We used the proton observations from the spacecraft (SC) 4 (Tango) since for this study it is necessary that particle instruments have been operating nearly continuously from 2001 through the present day. The L^* range we chose to study is L^*=5-9 as we introduced above.

Based on the observations by the Cluster Ion Spectrometry (CIS) using the time-of-flight ion Composition Distribution Function (CODIF) sensor (Reme et al., 2001), the low-energy component of the proton pressure (P_{CIS}) with the energy range of ~40 eV to 40 keV is calculated using the formula

$$P_{CIS}[nPa] = 1.381 \cdot 10^{-2} n[cm^{-3}] T[MK],$$

where n is the proton density, T is the proton temperature. The proton densities and temperatures can be found at CSA under the product C4_CP_CIS-CODIF_HS_H1_MOMENTS.

Based on the observations by the Research with Adaptive Particle Imaging Detector (RAPID) (Wilken et al., 2001), the high-energy part of the proton pressure (P_{RAPID}) with the energy range up to 4MeV is calculated using the formula (Daly & Kronberg, 2018; Kronberg et al., 2017):

$$P_{RAPID}[nPa] = 4\pi \frac{2}{3} 0.518 \cdot 10^{-8} \sqrt{m[amu]} \sum_{E} \Delta E[keV] \sqrt{E[keV]}\, j[cm^{-2} sr^{-1} s^{-1} keV^{-1}],$$

where m is the proton mass in atomic mass units (amu), j is the omnidirectional energetic proton intensity, ΔE is the width of the energy channel, and E is the effective energy. The geometric mean is used as an approximation for the effective energy (Kronberg & Daly, 2013). The omnidirectional energetic proton intensities can be found at CSA under the product Proton_Dif_flux__C4_CP_RAP_HSPCT. The first proton RAPID energy channel at 27.7–64.4 keV overlaps with the last CIS energy channel. Thus, in order to obtain a continuous spectrum, the first RAPID channel was truncated according using the method provided by Kronberg et al. (2022). The CIS and the RAPID instruments are well cross-calibrated for protons (Kronberg et al., 2010, 2022).

The steps of data processing are as follows. First, the proton data from both CIS and RAPID with original 4-second resolution was averaged over 1 minute in order to be consistent with the predictors related to the solar and geomagnetic activity which have the highest resolution of 1 minute. Then, the outliers were eliminated. The data points with calculated $P_{CIS} = 0$ were removed because both proton densities and temperatures were 0 at these points. When calculated $P_{CIS}$was above 100nPa, the data points were also removed. Because based on previous studies (De Michelis et al., 2013; Kronberg et al., 2017; Lui, 2003), proton pressures have never been over 100nPa even under disturbed geomagnetic

conditions. The large P_{CIS} at these points may be caused by the large proton densities (>100 cm^{-3}) due to the background contamination of CODIF. When calculating P_{RAPID}, we had to remove the data points with the flux of the second energy channel (75.3-92.2 keV) being 0. The slope of the energy spectra to derive the fluxes in the first truncated channel cannot be calculated in this case, leading to the inaccurate P_{RAPID}. Next, we added the P_{CIS} and P_{RAPID} with the same timestamp to get the total proton pressure. Finally, we chose the points with the position range of L^*=5-9 for our study. We also removed the points with the total proton pressures less than 0.1nPa because these values are not typical at these distances on the closed magnetic field lines. Most previous work showed that the values of the proton pressures were mostly above 0.1nPa in the inner magnetosphere (Kronberg et al., 2017; Lui, 2003). In addition, there were only 1580 points (0.5% of the total points) less than 0.1nPa. We obtained better model performance when dropping the values less than 0.1nPa. In this case, the model was more focused on predicting values above 0.1nPa, namely the typical values of proton pressures in the inner magnetosphere.

2. 2 Predictors

In this subsection, we divide predictors into two groups for description: related to the location in the space and related to the solar, solar wind, and geomagnetic activity. All the predictors are listed in **Table 1**.

2.2.1 Location in the Space

The predictors related to the location in the space include: L^* value, magnetic local time (MLT), and the position of Cluster in the Geocentric Solar Ecliptic (GSE) coordinate system (x_gse, y_gse and z_gse). The L^* values were taken from the Lstar_value__C4_CP_AUX_LSTAR dataset. MLTs can be found under the product Mag_Local_time__C4_JP_AUX_PMP. The sc_r_xyz_gse__C4_CP_AUX_POSGSE_1M dataset is the source of the position data. The distributions of the proton plasma pressures and the numbers of their samples in the GSE system are shown in **Figure 1**. The distributions of the proton plasma pressures and the number of their samples in the L^*-MLT coordinate are shown in **Figure 2**.

From **Figures 1d-f**, we can see that the number of samples is larger on the dayside and especially

at lower L^* shells. As we mentioned above, our L^* values are from "Lstar" product (A. G. Smirnov et al., 2020) which is calculated with Tsyganenko-89 magnetospheric model, T89 (Tsyganenko, 1989). Thus, one can understand that the Cluster trajectories can cover more samples in the dayside than in the nightside at the same L^*-shell range. This phenomenon can also be seen in **Figure 2b**, the numbers of samples in the L^*-MLT coordinate. However, this has no effects on the final results because the value in each bin is the median of at least hundreds of samples, which is enough to reduce the error and represent the median feature of the bin.

Figures 1a and **1b** show that the proton pressures are higher at the nightside than that at the dayside in the XY and XZ planes in the GSE coordinate system. Similarly, the same result can be observed in the L^*-MLT coordinate at L^*-shell > 5 in **Figure 2a**. **Figure 1a** also shows the dusk-dawn asymmetry at the dayside with higher proton pressures at the dusk side. The same higher pressures at the dusk side (+y side) are visible in **Figure 1c** (YZ plane). Likewise, we can note this dusk-dawn asymmetry at the dayside at lower L^* shells in **Figure 2a.** The reasons for these asymmetries will be discussed in **Section 5**.

In **Figures 3a-e**, the relations of mean proton pressures versus the predictors related to the location in the space are shown. **Figure 3a** shows a strong linear decrease of the proton plasma pressure with the L^* shell which is consistent with previous results as we introduced in **Section 1.** That is, no matter under quiet conditions (Lui & Hamilton, 1992) or disturbed conditions (De Michelis et al., 2013), the proton plasma pressure is monotonically decreasing with L^* shell when L^*=5-9. In **Figure 3b**, there is a peak of the proton pressure around midnight (0-2MLT) and the minimum of proton pressures is shown around 9 MLT. The proton pressure displays a peak at ~-6R_E and roughly linear decrease with XGSE coordinate in the distances between -6 R_E and 9 R_E in **Figure 3c**. This is also consistent with the day-night asymmetry with higher proton pressures at the nightside (-x side) in **Figures 1a** and **1b**. **Figure 3d** shows that the maximum of the proton pressure is around ±5 RE in YGSE direction. This YGSE dependency resembles that observed for the proton intensities in the near-Earth space in Kronberg et al. (2021). In **Figure 3e**, we can take ZGSE=-2R_E as the symmetry axis, the proton pressures on both sides decrease almost linearly with the direction away from the symmetry axis.

2.2.2 Solar, Solar Wind and Geomagnetic Activity

The predictors are also related to the observations of solar, solar wind, and geomagnetic parameters from the OMNI database (King & Papitashvili, 2005). The solar wind parameters we used include: the proton density, NpSW [cm^{-3}]; components of the velocity (VSW) in the GSE coordinates, VxSW_GSE, VySW_GSE and VzSW_GSE [$km \cdot s^{-1}$]; the proton temperature, Temp [K] and components of the IMF in the GSE coordinates, BimfxGSE, BimfyGSE and BimfzGSE [nT]; and the dynamic pressure, Pdyn [nPa]. The proton pressures increase with the solar wind velocity in the anti-sunward direction, V_x, in **Figure 3f.** The V_y and V_z components are associated with increase of the proton pressures when they deviate from the Earth-Sun direction (V_x) within a certain range (<±100 km/s), as shown in **Figure 3g**. When any component of the interplanetary magnetic field (IMF) becomes stronger, no matter in positive or negative directions, it may lead to the increased proton pressures (see **Figure 3h**). These dependencies of the proton pressures on the solar wind velocity components and the IMF components also resemble those observed for the proton intensities in the near-Earth space in Kronberg et al. (2021). In order to show the relationships more clearly, we plot the figures in the logarithmic scale in **Figures 3i-k**. In **Figure 3i**, when the solar wind density is greater than 2 cm^{-3}, the proton pressures generally increase with it. There is an approximate linear relationship between the proton pressures and the logarithm of the solar wind dynamic pressures when the solar wind dynamic pressure is greater than 1nPa (see **Figure 3j**). In **Figure 3k**, a trend of increase of the proton pressures with the solar wind temperature is visible. The 10.7 cm solar radio flux ($F_{10.7}$) is one of the most widely used indices of solar irradiance (Tapping, 2013). Kistler and Mouikis (2016) showed that $F_{10.7}$ have an impact on the proton flux at L=6-7. In **Figure 3l**, the solar irradiance is non-linearly related to the proton pressure, but a general trend of decrease of the proton pressures with the F10.7 is visible.

For the parameters related to the geomagnetic activity, we used AE index and SYM-H index. The auroral electrojet index (AE index) provides a global, quantitative measure of auroral zone magnetic activity within the auroral oval. The proton pressures are related roughly linear up to ~ 600nT with the logarithm of the AE index in **Figure 3m**. SYM-H index describes symmetric horizontal component disturbances of the geomagnetic field at the equatorial regions (Iyemori et al., 2010). SYM-H index,

shows non-linear relation with proton pressures, see **Figure 3n**. The histograms of the number of samples of all these predictors are shown in **Figure 12** in **Appendix**.

2.2.3 Cross-correlations between Proton Pressures and Predictors

In **Figure 4,** we show the Pearson linear correlations between proton pressures and the predictors. This measurement can only reflect a linear correlation of variables, and ignore other types of correlations. The range of this correlation value is from -1 to 1. Values close to -1/1 mean perfect linear anticorrelation/correlation and values equal to 0 mean there is no linear dependency between the variables. The proton pressures are well anticorrelated with the L^* shell (-0.51), in agreement with **Figure 3a**. For the XGSE and ZGSE location of observation, they also show some anticorrelation with the proton pressures, -0.21 and -0.14, respectively. From the OMNI parameters, the proton pressures are best linearly correlated with the solar wind dynamic pressure, 0.23. The AE-index also shows some correlation with proton pressures (0.16), the same as the result in **Figure 3**.

3. Methodology

3.1 Data Split

After the data processing as we mentioned in **Section 2.1**, the full dataset we used comprises in total 336481 measurements from 2001-02-04 12:31:00 UT to 2018-02-18 00:02:00 UT. We split the dataset into a training set (80%) and a test set (20%). To prevent test leakage, we split the data by a time point with the original order preserved (Camporeale, 2019; Kronberg et al., 2021). The test set is only for the testing of the model. After the model training has been completed, no further changes to the model can be made. We utilize the training set to train and optimize the model hyperparameters. The sizes and periods of data subsets after splitting are listed in **Table 2**.

Machine Learning algorithms don't perform well when the input numerical features have very different scales. Thus, we need to normalize the features in order to get all the features to have the same scale (Géron, 2019). We normalized the features by QuantileTransformer in sklearn (Pedregosa et al., 2011). QuantileTransformer provides a non-parametric transformation to map the data to a uniform distribution with values between 0 and 1. This transformation smooths out unusual distributions and is

less influenced by outliers than other methods (Pedregosa et al., 2011).

3.2 Machine Learning Models for Proton Pressures

Our study is one of the typical supervised learning tasks, called regression. We have applied various kinds of regression ML models in order to select the best one based on their validation performance. We note that most of the relations between the proton pressures and the predictors are not perfectly linear, as shown in **Figure 3.** In addition, the ensemble of the predictions of a group of predictors (such as regressors) will often give better predictions than with the best individual predictor (Géron, 2019). Thus, we not only tried linear regression models, but also ensemble regression models.

We have examined the following linear models in sklearn.linear_model and sklearn.svm (Pedregosa et al., 2011): (1) Ridge Regression, namely linear least squares with l2 regularization (Hoerl & Kennard, 1970); (2) Least Angle Regression (LARS) (Efron et al., 2004); (3) Linear Support Vector Regression (LinearSVR) (Cortes & Vapnik, 1995).

We also consider the Decision Trees Regression (Breiman et al., 1984) in sklearn.tree and the tree-based ensemble models: (1) Random Forest Regression (Ho, 1995); (2) Extra Trees Regression, namely extremely randomized trees (Geurts et al., 2006); (3) AdaBoost Regression (Freund & Schapire, 1997); (4) Gradient Boosting Regression (Friedman, 2001); (5) Histogram-Based Gradient Boosting Regression ((1)-(5) are all from sklearn.ensemble); (6) Light Gradient Boosting Machines (LGBM) (Ke et al., 2017) in LightGBM library.

In order to evaluate different models' performances, we focus on the Spearman correlation. The Spearman correlations between the model results and the observations are listed in **Table 3.** Pearson correlation only assesses linear relationships as we discussed in **Figure 4**, while Spearman correlation assesses monotonic relationships (whether linear or not). The values of the Spearman correlation vary between -1 and 1. Correlations of -1 or +1 imply an exact monotonic relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases. Values close to 0 means no monotonically correlation. In **Table 3**, we can note that the Extra-Trees Regressor has shown the best predicting performance on the both sets. Although the Decision Tree Regressor also

has a perfect performance on the train/validation set, it seems to be more inclined to the overfitting (the difference between the scores for the train/validation set and test set are larger). In addition, note that a gap between model performance on training and test data is often observed for complex models (Kronberg et al., 2021). Extra-Trees Regressor fits a large amount of randomized decision trees on the training dataset and uses the mean to improve the predictive accuracy and control overfitting. It has two main differences with other tree-based ensemble methods: (1) it splits nodes by choosing cut-points fully at random. That is, besides searching for the best feature among a random subset of features, like the regular Random Forests, it also utilizes random thresholds for each feature rather than searching for the best possible thresholds. (2) it uses the whole learning sample (rather than a bootstrap replica) to grow the trees (Géron, 2019; Geurts et al., 2006). These characteristics can result a lower variance and a faster training speed (compared with regular Random Forests). Thus, we decided to use Extra-Trees Regressor.

3.3 Training the selected model

We trained the model using the K-Fold cross-validation (CV) function (from sklearn.model_selection.KFold). This method is widely used in previous work (e.g., Kronberg et al., 2020; A. Smirnov et al., 2020). Our training data are divided into K subsets (folds) which are roughly the same size. In our case, K =5. Then each fold is used once as a validation while the 4 remaining folds form the train set. In this way, splitting process is repeated 5 times and results in five arrays of evaluation scores. Cross-validation allows one to get not only an estimate of the performance of the model, but also a measure of how precise this estimate is (Géron, 2019).

In order to determine the best hyperparameters, the parameters are optimized by grid-search over a parameter grid, using GridSearchCV (from sklearn.model_selection.GridSearchCV). To evaluate the performance of the training and validation during the cross-validation for different parameters, we use four assessment metrics: Spearman correlation, mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2). The best scores for MSE and MSE are close to 0. R^2 is a number between 0 and 1 that measures how well a statistical model predicts an outcome. That is, comparing the case of using the model for prediction with the case of only using the mean prediction, to see how much the performance of the model has been improved. In the perfect case, R^2 is equal to 1. The resulting

hyperparameters values, as well as their search ranges, are given in **Table 4**. The performances of the model for the train/validation data set are mostly consistent between different metrics. n_estimators controls the number of trees in the forest. It will be underfitting when n_estimators is too small, while it will be overfitting when n_estimators is too large. max_depth is the maximum depth of the tree. If max_depth=None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. min_samples_split is the minimum number of samples required to split an internal node. We use the default value 2 for min_samples_split. Another important parameter we optimized is the minimum number of data points in leaf (min_data_in_leaf), which has a regularization effect and stops the model from learning the noise.

4. Results

4.1 Test the model

The final scores are the average performances of the model for the train/validation data set, see **Table 5**. The values of the Spearman correlation coefficient are very close for the test data (0.68) and the average validation (0.71). The mean squared (MSE) and absolute errors (MAE) also yield almost identical values for validation and test sets. This means that our model is not overfitting and successfully learns relationships between the input parameters and the resulting proton pressures and generalizes well onto the unseen data.

The Spearman correlation between the observed and predicted data is about 68% for the test set. This value is reasonable considering the complex dynamics of the energetic protons in the inner magnetosphere. In addition, when we use the Spearman correlation to evaluate our model, there is a null hypothesis states that the predictions are uncorrelated to the observations (evaluated by p value). We obtain p=0. In other words, we can reject the null hypothesis, namely the model predictions are correlated to the observations. Thus, our model results are reliable and can learn the overall trend in the proton pressures.

4.2 Visualized Results

Figure 5a shows the distribution of the observed proton plasma pressures versus the predicted values from the training set, while **Figure 6a** represents the test set distribution. Observed and predicted

data for the training and test data sets agree relatively well. The diagonal shows the one−to−one ratio between the observed and predicted pressures. The data is mainly concentrated along the black dashed line, corresponding to a good correlation. The histograms in **Figures 5a** and **6a** represent the predicted or observed data points that fall into each corresponding bin. **Figures 5b** and **6b** provides the histogram of model residuals. From these figures, we can note that for both the training set and the test set, our model has very low bias. Most of the model residuals are within the range of ±0.5, namely the ratios of observations and predictions are valued in ~0.3-~3 ($10^{-0.5}$-$10^{0.5}$). Therefore, we can conclude that our final model predicts the proton pressures at L^*=5-9 well since it has low bias and can capture the general trends represented in the data.

In **Figure 7**, we show a qualitative example of the model's predictions within the 6-hour time interval on 2017-07-04 (in the test set) under quiet geomagnetic conditions (SYM_H>-1). The model almost predicts the same proton plasma pressures with the observations in **Figure 7c**. **Figure 8** shows another example on 2017-09-28 (in the test set) which demonstrates the model performance during the main phase of a magnetic storm with the SYM-H index dropping down to ~-68nT. We can note in the panel c that the predictions are almost always lower than the observations under disturbed geomagnetic conditions. This is because the main phases of the magnetic storms are rather rare events in our dataset. Our model is not developed specifically for the prediction of the proton pressures under disturbed geomagnetic conditions. The ML model of soft proton intensities by Kronberg et al. (2021) also has better prediction efficiency under quiet geomagnetic conditions.

We also plot the distributions of the predicted proton plasma pressures in L*-MLT coordinates under quiet (**Figure 7d**) and disturbed geomagnetic conditions (**Figure 8d**). The input predictors are the median values of the parameters over the time period (except the location parameters: X/Y/Z_GSE, L* value and MLT) in the purple region in **Figures 7** and **8**. The details of the input predictors are listed in **Table 6** in the **Appendix**. For calculations of L* values, it is necessary to specify the satellite position, magnetic field model and geomagnetic conditions (by 'get_Lstar' function in IRBEM library). The initial position range we give in each direction is [-11, 11] R_E in the GSE system, which is consistent with the X/Y/Z_GSE range of our observation dataset. Tsyganenko-89 magnetospheric model, T89 (Tsyganenko,

1989) with Kp index as an input is employed, as Smirnov et al. (2020) did. The Kp index for quiet geomagnetic times is set as 0, while for disturbed geomagnetic times is set as 4. The time moments indicated by the red dotted lines in **Figures 7** and **8** are specified for the calculations of MLTs. All of the calculations are performed using the spacepy.irbempy library ('get_Lstar' function). Finally, we select the data points that fit the range of L*=5-9 from the output results for plotting **Figures 7d** and **8d**.

We can note that there are no results when L*>~8 under disturbed geomagnetic conditions in **Figure 8d**. L* is the property of a stably trapped particle. A pseudo-trapped particle (particles that will leave the magnetosphere before completing a 180° drift) that drifts into the magnetopause (magnetopause shadowing) or into the tail (tail-shadowing) does not have an L*-value (Roederer, 1967; Roederer & Lejosne, 2018). The particles are easier become pseudo-trapped particles and to be lost during disturbed times (Roederer & Lejosne, 2018) since the magnetosphere is compressed based on T89 model (Tsyganenko, 1989).

From the predictions in **Figure 7d** and **8d**, we can note the dusk-dawn asymmetry at the dayside with higher proton pressures at the dusk side and the day-night asymmetry with higher proton pressures at the night side**.** In regard of the day-night asymmetry under quiet geomagnetic conditions in **Figure 7d** we consider the higher L^* shells (L^*>6). In addition, we can note that the proton pressures at the nightside under disturbed geomagnetic conditions seem to be higher than that under quiet geomagnetic conditions. The proton pressures at the afternoon sector (12-18MLT) under quiet geomagnetic conditions seem to be higher than that under disturbed geomagnetic conditions. A quantitative analysis of this phenomenon and the reasons for these asymmetries will be discussed in **Section 5**.

The plotting process of **Figure 9b** is the same as that of **Figures 7d** and **8d**, except that the input predictors are the median values of the parameters over the whole dataset time. The details of the input predictors are listed in **Table 6** in the **Appendix**. The Kp index is set as 0. The **Figure 9a** and **Figure 2a** is the same figure, namely the distribution of the observed H^+ plasma pressures over the whole dataset time. In **Figure 9b**, we can note the dusk-dawn asymmetry at the dayside and the day-night asymmetry, just as the observations in **Figure 9a**. By comparing the results of observations (**Figure 9a**) and predictions by our model (**Figure 9b**) based on the whole dataset, we can conclude that our model can

reproduce the overall characteristics of the distributions of observed proton plasma pressures in the range of L*=5-9.

4.3 Feature Importance

One of the advantages of the tree-based machine learning models is that they make it easy to measure the relative importance of each feature (Géron, 2019). Scores are automatically computed for each feature after training. The values of all the feature importances sum to 1. This process is called feature importance. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature, also known as the Gini importance. The higher the value, the more important the feature is. **Figure 10** shows the feature importance for each input variable. The black horizontal lines represent confidence intervals at 95% confidence level. The parameters related to the location show significantly higher importance than parameters related to solar, solar wind, and geomagnetic activity. From those, on average, the strongest dependence is seen for L^* shell. This is also consistent with the results in **Figures 3a** and **4**. The least important location parameter is y_gse. From the other parameters, the solar wind dynamic pressure is the most important parameter for predicting the proton plasma pressures. Based on the observations of multiple satellites, Stepanova et al. (2019) also found that the plasma pressure inside the magnetosphere is mainly controlled by the solar wind dynamic pressure, which can be related to the pressure balance at the magnetospheric flanks.

5. Discussions

Based on the observations and predictions by our model under quiet (**Figure 7**) and disturbed (**Figure 8**) geomagnetic conditions, we note that no matter under quiet or disturbed geomagnetic conditions, both the dusk-dawn asymmetry at the dayside and the day-night asymmetry occur. The persistent dusk-dawn asymmetry with higher proton pressures at the duskside may be related to the dawn-dusk asymmetry of the proton distribution in the plasma sheet. Based on 7-years observations of energetic protons >274 keV by RAPID instrument, Kronberg et al. (2015b) showed the dawn-dusk asymmetries of proton intensities in the plasma sheet at near-Earth nightside under both quiet and disturbed geomagnetic conditions. They also explained two general effects which can lead to this kind of dawn-dusk asymmetry. In addition, other previous work also reported dawn-dusk asymmetries in the

plasma sheet of energetic particles intensities with different energy ranges (Meng et al., 1981; Sarafopoulos et al., 2001; Kistler & Mouikis, 2016).

The day-night asymmetry is easy to understand because ions are injected into the inner magnetosphere through the plasma sheet at the nightside, especially during the magnetic storms or substorms (Kistler et al., 1992). Gabrielse et al. (2014) indicated that injection occurrence rates increase with the geomagnetic activity. This may also be the reason for the higher proton pressures at the nightside under disturbed geomagnetic conditions as we shown above. For further quantitative analysis of this phenomenon, **Figure 11** was plotted to investigate the difference between the proton plasma pressure under disturbed and quiet geomagnetic conditions. The red colors are positive values which means that the proton pressures are higher under disturbed geomagnetic conditions than under quiet geomagnetic conditions, while the blue colors are the opposite. The difference was calculated by the predictions of our model (**Figures 7d** and **8d**). In **Figure 11**, we can note that the proton pressures at the nightside under disturbed geomagnetic conditions are clearly higher than that under quiet geomagnetic conditions. In addition, more red bins are seen at the lower L* shells (L*=5-6) than at the higher L* shells (L*>6) at the nightside, which means that there are higher increases in the plasma pressures at the lower L* shells (L*=5-6) during disturbed times. This is consistent with the results of Figure 6b in Gabrielse et al. (2014). Namely, injections more frequently reach lower L-shells with increased geomagnetic activity. This can also be the reason why the day-night asymmetry under quiet geomagnetic conditions in **Figure 7d** mainly concentrates on the higher L^* shells (L^*>6).

In addition, the proton pressures at 12-18MLT sector (afternoonside) under disturbed geomagnetic conditions are clearly lower than that under quiet geomagnetic conditions. This may be related to the outflow of energetic ions through the magnetopause in the dayside under disturbed geomagnetic conditions. Keika et al. (2005) showed that the outflowing energy flux is higher on the afternoonside than that on the morningside during the main phase of magnetic storms, which may lead to the lower proton pressures on the afternoonside under disturbed geomagnetic conditions. Estimation of a comparison between the losses at the magnetopause and the difference between the proton plasma pressure on the afternoonside under disturbed and quiet geomagnetic conditions requires further studies.

Thus, we can deduce that the patterns of the asymmetries may change with the geomagnetic conditions. However, a more detailed calculation of the asymmetry index (e.g., Luo et al.,2017) separately for the quiet and the disturbed time is beyond the scope of this paper and will be further studied in the future.

6. Conclusions

In this study, based on 17-year data from both CIS and RAPID instruments onboard the Cluster mission, we derive a machine-learning-based model for predicting proton pressures at energies from ~40eV to 4MeV at the outer part of the 3D inner magnetosphere (L^*=5-9). The results demonstrate that the Extra-Trees Regressor shows the best predicting performance. The Spearman correlation between the observed and predicted data is about 68% despite the complex dynamics of the energetic protons in the magnetosphere. The most important parameter for predicting proton pressures in our model is the L^* shell, related to the location. The most important predictor of solar, solar wind, and geomagnetic activity is the solar wind dynamic pressure. The model results are in general agreement with the previous studies (De Michelis et al., 2013; Lui & Hamilton, 1992; Stepanova et al., 2019). In addition, we use the model prediction to compare and explain the distributions of the proton plasma pressures under different geomagnetic conditions. Moreover, as we discussed in the introduction, our results can be used in the simulations of the inner magnetosphere (e.g., HEIDI model) or reconstructing the 3-D electric current system. It can also provide valuable guidance to the space weather forecast.

Further directions for the present study include, first, incorporating oxygen ions data into the model in order to predict the complete 3D distribution of ion plasma pressures in the outer part of the inner magnetosphere. Second, a machine-learning-based model for predicting the 3-D ion pressures in the inner part of the inner magnetosphere (L^*=2-5). This aim can be achieved by using data from other missions, such as Van Allen Probes. The results of the model for the ion pressures in the inner part of the magnetosphere will be compared with the results of this model. In addition, we can combine these two models together to predict the 3-D ion pressures in the complete inner magnetosphere (L^*=2-9).

Acknowledgement

The authors are thankful to the Cluster Science Archive team (https://csa.esac.esa.int) for providing the

data. We acknowledge the use of NASA/GSFC's Space Physics Data Facility's OMNIWeb service and OMNI data. We acknowledge the use of the IRBEM library (V4.3), the latest version of which can be found at https://doi.org/10.5281/zenodo.6867552. This work is supported by the National Natural Science Foundation of China (41874197). S.Y.Li is also supported by the China Scholarship Council (award to S.Y. Li for 1 year study abroad at Ludwig Maximilians University Munich). EK is supported by German Research Foundation (DFG) under number KR 4375/2-1 within SPP "Dynamic Earth".

References

Antonova, E. E. (2004). Magnetostatic Equilibrium and Current Systems in the Earth's Magnetosphere. Streamers, Slow Solar Wind, and the Dynamics of the Magnetosphere, 33(5), 752-760. http://doi.org/10.1016/S0273-1177(03)00636-7

Antonova, E. E., Kirpichev, I. P., & Stepanova, M. V. (2014). Plasma Pressure Distribution in the Surrounding the Earth Plasma Ring and Its Role in the Magnetospheric Dynamics. Journal of Atmospheric and Solar-Terrestrial Physics, 115, 32-40. http://doi.org/10.1016/j.jastp.2013.12.005

Brandt, P. C. s., Roelof, E. C., Ohtani, S., Mitchell, D. G., & Anderson, B. (2004). Image/Hena: Pressure and Current Distributions During the 1 October 2002 Storm. Advances in Space Research, 33(5), 719-722. http://doi.org/10.1016/s0273-1177(03)00633-1

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). Classification And Regression Trees (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Camporeale, E. (2019). The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. Space Weather-the International Journal of Research and Applications, 17(8), 1166-1207. http://doi.org/10.1029/2018sw002061

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine learning, 20(3), 273-297. http://doi.org/Doi 10.1007/Bf00994018

Daly, P., & Kronberg, E. (2018). User Guide to the Rapid Measurements in the Cluster Science Archive (CSA): Version 5.2, Tech. Rep. CAA-EST-UG-RAP, European Space Agency, Paris.

De Michelis, P., Daglis, I. A., & Consolini, G. (2013). Average Terrestrial Ring Current Derived from Ampte/Cce-Chem Measurements. Journal of Geophysical Research: Space Physics, 102(A7), 14103-14111. http://doi.org/10.1029/96ja03743

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. The Annals of statistics, 32(2), 407-499.

Escoubet, C., Schmidt, R., & Goldstein, M. (1997). Cluster-Science and Mission Overview. The cluster and phoenix missions, 11-32.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. Journal of computer and system sciences, 55(1), 119-139. http://doi.org/DOI 10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of statistics, 29(5), 1189-1232. http://doi.org/DOI 10.1214/aos/1013203451

Gabrielse, C., V. Angelopoulos, A. Runov, and D. L. Turner (2014), Statistical characteristics of particle injections throughout the equatorial magnetotail, J. Geophys. Res. Space Physics, 119,2512–2535, doi:10.1002/2013JA019638.

Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems: " O'Reilly Media, Inc.".

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely Randomized Trees. Machine learning, 63(1), 3-42. http://doi.org/10.1007/s10994-006-6226-1

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression - Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55-&. http://doi.org/Doi 10.1080/00401706.1970.10488634

Ho, T. K., "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

Ilie, R., M. W. Liemohn, G. Toth, and R. M. Skoug (2012), Kinetic model of the inner magnetosphere with arbitrary magnetic field, J. Geophys. Res., 117, A04208, doi:10.1029/2011JA017189.

Iyemori, T., Takeda, M., Nose, M., Odagi, Y., & Toh, H. (2010). Mid-Latitude Geomagnetic Indices "Asy" and "Sym" for 2009 (Provisional). Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University, Japan.

Ke, G. L., Meng, Q., Finley, T., Wang, T. F., Chen, W., Ma, W. D., et al. (2017). Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (Nips 2017), 30. Retrieved from <Go to ISI>://WOS:000452649403021

Keika, K., Nose, M., Ohtani, S., Takahashi, K., Christon, S. P., & McEntire, R. W. (2005). Outflow of Energetic Ions from the Magnetosphere and Its Contribution to the Decay of the Storm Time Ring Current. Journal of Geophysical Research-Space Physics, 110, A09210, doi:10.1029/2004JA010970.

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.

Kistler, L. M., Mobius, E., Baumjohann, W., Paschmann, G., & Hamilton, D. C. (1992). Pressure Changes in the Plasma Sheet During Substorm Injections. Journal of Geophysical Research-Space Physics, 97(A3), 2973-2983. http://doi.org/Doi 10.1029/91ja02802

Kistler, L. M., & Mouikis, C. G. (2016). The Inner Magnetosphere Ion Composition and Local Time Distribution over a Solar Cycle. Journal of Geophysical Research: Space Physics, 121(3), 2009-2032. http://doi.org/10.1002/2015ja021883

Kronberg, E. A., & Daly, P. W. (2013). Spectral Analysis for Wide Energy Channels. Geoscientific Instrumentation, Methods and Data Systems, 2(2), 257-261.

Kronberg, E. A., Daly, P. W., Dandouras, I., Haaland, S., & Georgescu, E. (2010). Generation and Validation of Ion Energy Spectra Based on Cluster Rapid and Cis Measurements. In The Cluster Active Archive (pp. 301-306): Springer.

Kronberg, E. A., Daly, P. W., & Vilenius, E. (2022). Calibration Report of the RAPID Measurements in the Cluster Science Archive (CSA), technical report. European Space Agency.

Kronberg, E. A., Gastaldello, F., Haaland, S., Smirnov, A., Berrendorf, M., Ghizzardi, S., et al. (2020). Prediction and Understanding of Soft-Proton Contamination in Xmm-Newton: A Machine Learning Approach. The Astrophysical Journal, 903(2), 89. http://doi.org/10.3847/1538-4357/abbb8f.

Kronberg, E. A., Grigorenko, E., Haaland, S., Daly, P. W., Delcourt, D. C., Luo, H., et al. (2015). Distribution of Energetic Oxygen and Hydrogen in the near‐Earth Plasma Sheet. Journal of Geophysical Research: Space Physics, 120(5), 3415-3431.

Kronberg, E. A., Hannan, T., Huthmacher, J., Münzer, M., Peste, F., Zhou, Z., et al. (2021). Prediction of Soft Proton Intensities in the near-Earth Space Using Machine Learning. The Astrophysical Journal, 921(1), 76.

Kronberg, E. A., Welling, D., Kistler, L. M., Mouikis, C., Daly, P. W., Grigorenko, E. E., et al. (2017). Contribution of Energetic and Heavy Ions to the Plasma Pressure: The 27 September to 3 October 2002 Storm. Journal of Geophysical Research: Space Physics, 122(9), 9427-9439. http://doi.org/10.1002/2017ja024215

Laakso, H., Perry, C., McCaffrey, S., Herment, D., Allen, A., Harvey, C., et al. (2010). Cluster Active Archive: Overview. The cluster active archive, 3-37.

Lui, A. T. Y. (2003). Inner Magnetospheric Plasma Pressure Distribution and Its Local Time Asymmetry. Geophysical Research Letters, 30(16).

Lui, A. T. Y., & Hamilton, D. C. (1992). Radial Profiles of Quiet Time Magnetospheric Parameters. Journal of Geophysical Research, 97(A12), 19325. http://doi.org/10.1029/92ja01539

Luo, H., Kronberg, E. A., Nykyri, K., Trattner, K. J., Daly, P. W., Chen, G. X., et al. (2017). Imf Dependence of Energetic Oxygen and Hydrogen Ion Distributions in the near‐Earth Magnetosphere. Journal of Geophysical Research: Space Physics, 122(5), 5168-5180.

http://doi.org/10.1002/2016ja023471
Meng, C. I., Lui, A., Krimigis, S., Ismail, S., & Williams, D. (1981). Spatial Distribution of Energetic Particles in the Distant Magnetotail. Journal of Geophysical Research: Space Physics, 86(A7), 5682-5700.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python. the Journal of machine Learning research, 12, 2825-2830.
Reme, H., Aoustin, C., Bosqued, J., Dandouras, I., Lavraud, B., Sauvaud, J., et al. (2001). First Multispacecraft Ion Measurements in and near the Earth's Magnetosphere with the Identical Cluster Ion Spectrometry (CIS) Experiment. Ann. Geophys., 19, 1303–1354, https://doi.org/10.5194/angeo-19-1303-2001, 2001.
Roederer, J. G. (1967). On the Adiabatic Motion of Energetic Particles in a Model Magnetosphere. Journal of Geophysical Research, 72(3), 981-992.
Roederer, J. G., & Lejosne, S. (2018). Coordinates for Representing Radiation Belt Particle Flux. Journal of Geophysical Research: Space Physics, 123(2), 1381-1387. http://doi.org/10.1002/2017ja025053
Sarafopoulos, D., Sidiropoulos, N., Sarris, E., Lutsenko, V., & Kudela, K. (2001). The Dawn-Dusk Plasma Sheet Asymmetry of Energetic Particles: An Interball Perspective. Journal of Geophysical Research: Space Physics, 106(A7), 13053-13065.
Sergeev, V. A., Pulkkinen, T. I., Pellinen, R. J., & Tsyganenko, N. A. (1994). Hybrid State of the Tail Magnetic Configuration During Steady Convection Events. Journal of Geophysical Research, 99(A12), 23571. http://doi.org/10.1029/94ja01980
Smirnov, A., Berrendorf, M., Shprits, Y., Kronberg, E. A., Allison, H. J., Aseev, N. A., et al. (2020). Medium Energy Electron Flux in Earth's Outer Radiation Belt (Merlin): A Machine Learning Model. Space Weather, 18(11), e2020SW002532.
Smirnov, A. G., Kronberg, E. A., Daly, P. W., Aseev, N. A., Shprits, Y. Y., & Kellerman, A. C. (2020). Adiabatic Invariants Calculations for Cluster Mission: A Long-Term Product for Radiation Belts Studies. Journal of Geophysical Research: Space Physics, 125(2), e2019JA027576.
Stepanova, M., Antonova, E. E., & Bosqued, J. M. (2008). Radial Distribution of the Inner Magnetosphere Plasma Pressure Using Low-Altitude Satellite Data During Geomagnetic Storm: The March 1–8, 1982 Event. Advances in Space Research, 41(10), 1658-1665. http://doi.org/10.1016/j.asr.2007.06.002
Stepanova, M., Antonova, E. E., Moya, P. S., Pinto, V. A., & Valdivia, J. A. (2019). Multisatellite Analysis of Plasma Pressure in the Inner Magnetosphere During the 1 June 2013 Geomagnetic Storm. Journal of Geophysical Research: Space Physics, 124(2), 1187-1202. http://doi.org/10.1029/2018ja025965
Stephens, G. K., Sitnov, M. I., Kissinger, J., Tsyganenko, N. A., McPherron, R. L., Korth, H., & Anderson, B. J. (2013). Empirical Reconstruction of Storm Time Steady Magnetospheric Convection Events. Journal of Geophysical Research: Space Physics, 118(10), 6434-6456. http://doi.org/10.1002/jgra.50592
Tapping, K. (2013). The 10.7 Cm Solar Radio Flux (F10. 7). Space Weather, 11(7), 394-406.
Tsyganenko, N. A. (1989). A Magnetospheric Magnetic Field Model with a Warped Tail Current Sheet. Planetary and Space Science, 37(1), 5-20.
Wing, S., & Newell, P. T. (1998). Central Plasma Sheet Ion Properties as Inferred from Ionospheric Observations. Journal of Geophysical Research: Space Physics, 103(A4), 6785-6800. http://doi.org/10.1029/97ja02994
Wilken, B., Daly, P., Mall, U., Aarsnes, K., Baker, D., Belian, R., et al. (2001). First Results from the Rapid Imaging Energetic Particle Spectrometer on Board Cluster. Ann. Geophys., 19, 1355–1366, https://doi.org/10.5194/angeo-19-1355-2001, 2001.

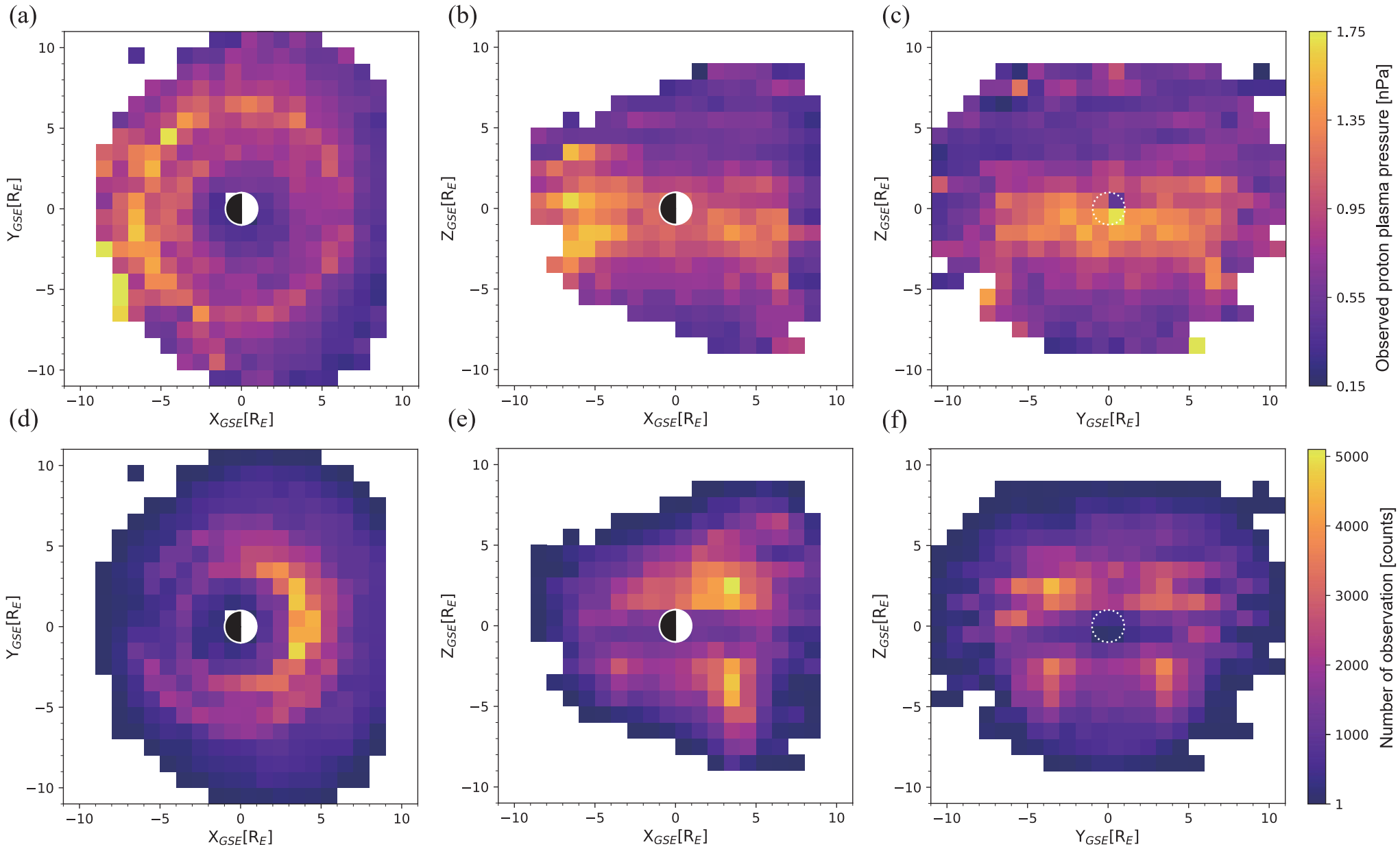

Figure 1. (a-c) Distributions of the observed H^+ plasma pressures by SC4 from February 2001 to February 2018 in the GSE coordinate system. (d-f) Distributions of the number of measurements corresponding to (a-c). Resolution (bin size) is 1 Re.

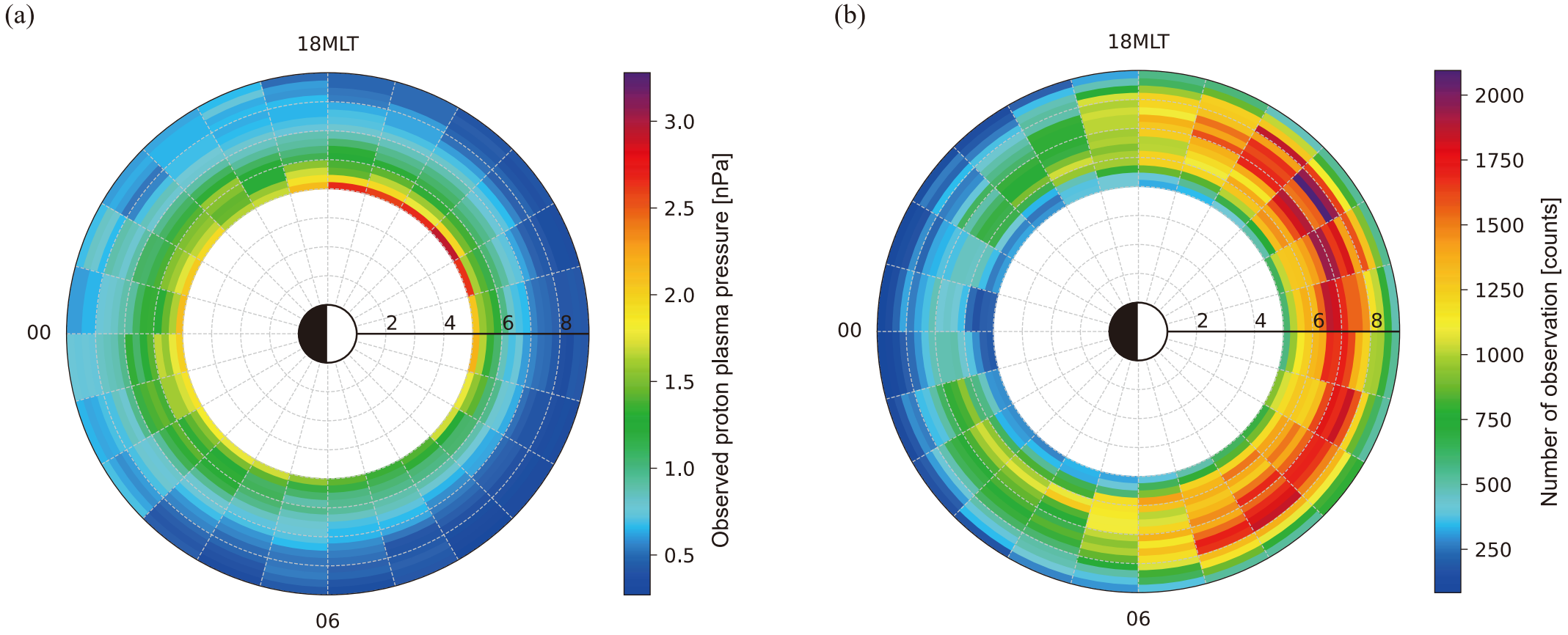

Figure2. Distributions of the observed H^+ plasma pressures (left) and the number of measurements (right) by SC4 from February 2001 to February 2018 in the L*-MLT coordinate. The L* resolution (bin size) is 0.25. MLT resolution (bin size) is 1.

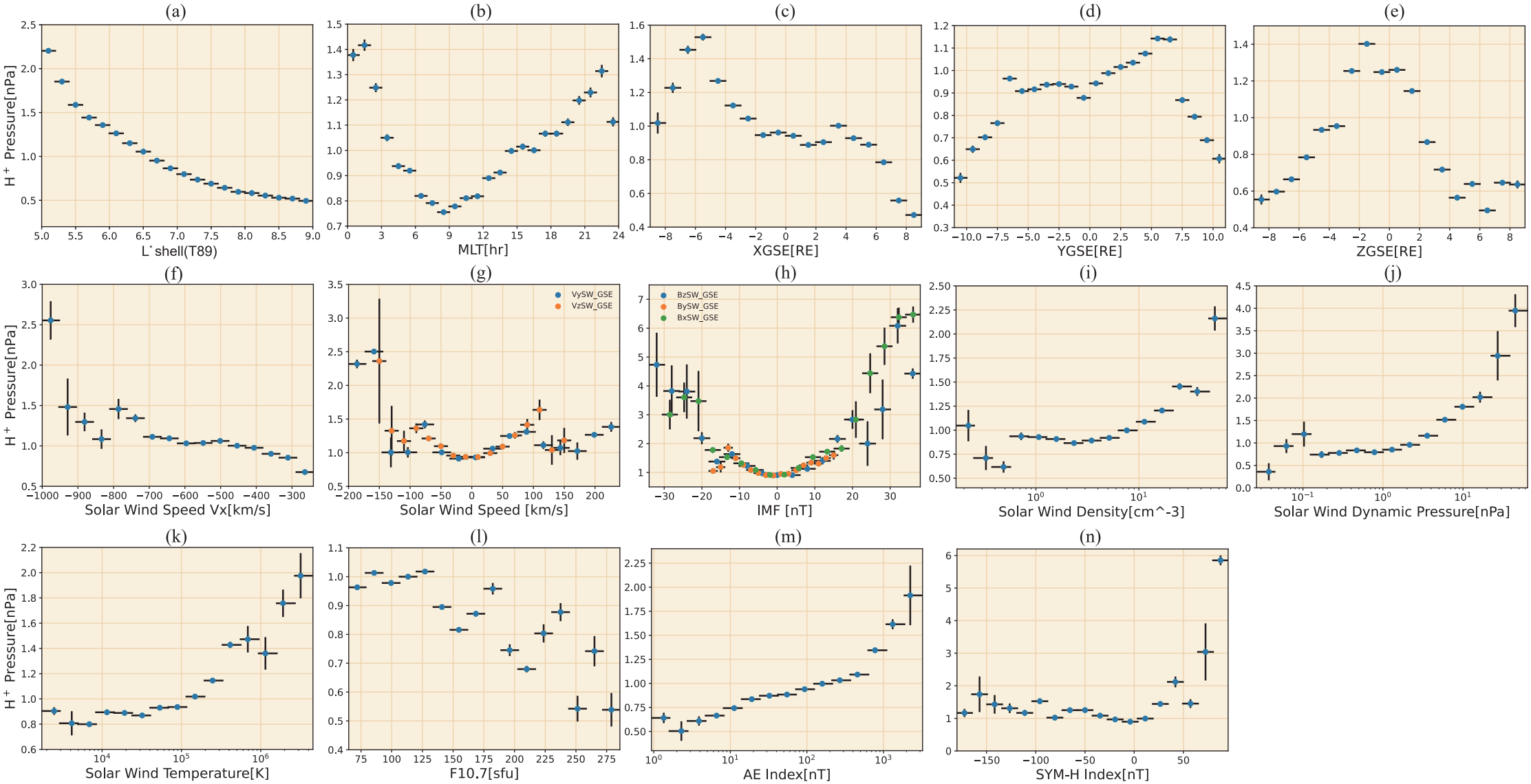

Figure3. Relations of mean H^+ plasma pressure and (a)-(e) L^* shell, MLT, XGSE, YGSE, and ZGSE, respectively; (f)-(g) the solar wind Vx ,Vy and Vz components; (h) IMF components in GSE;(i)-(k) solar wind density, dynamic pressure and temperature, respectively; (l) F10.7 parameter; (m) AE index and (n) SYM-H index. Vertical lines represent confidence intervals at 95% confidence level.

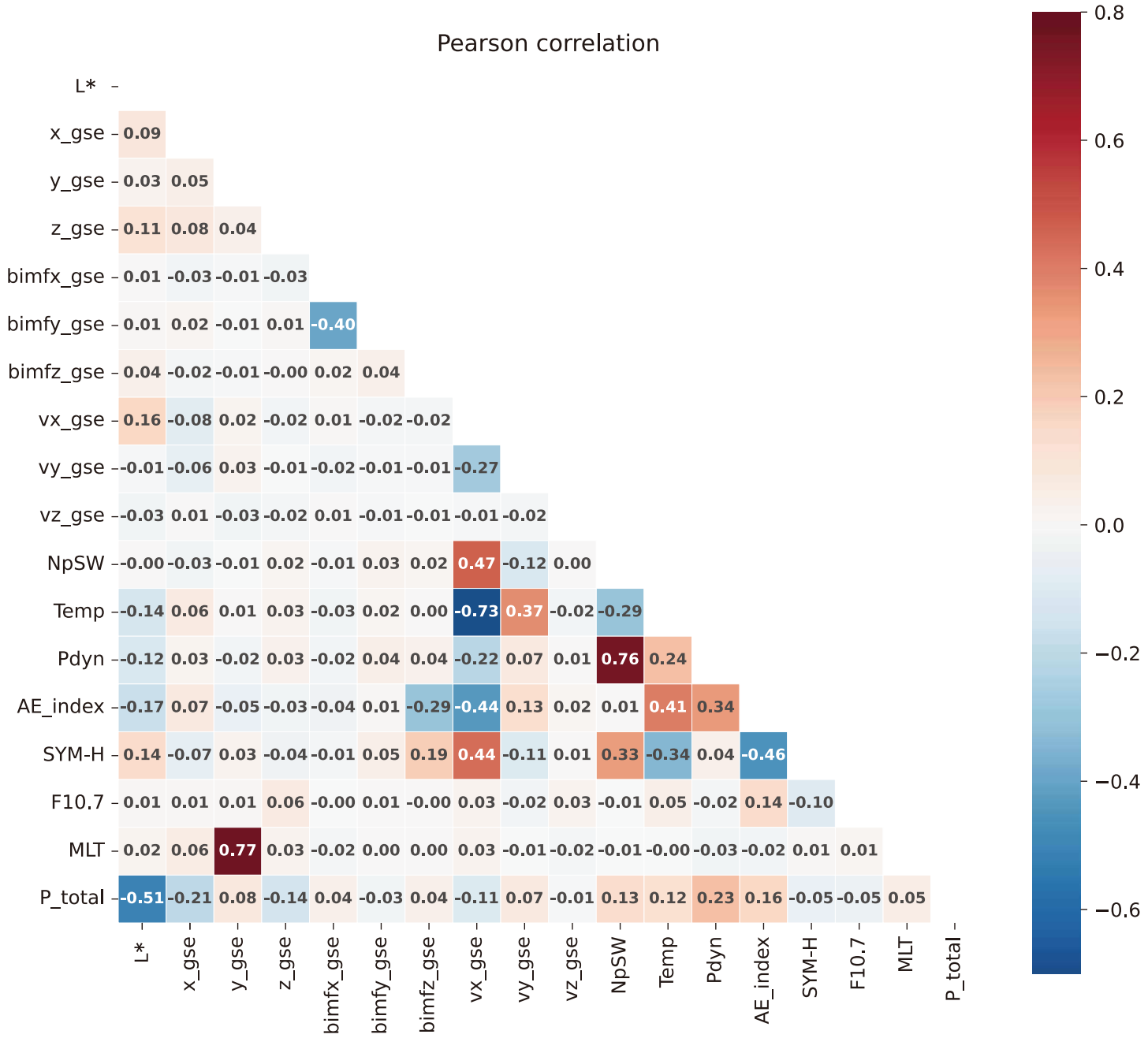

Figure4. Pearson correlation matrix between input parameters and H^+ plasma pressure.

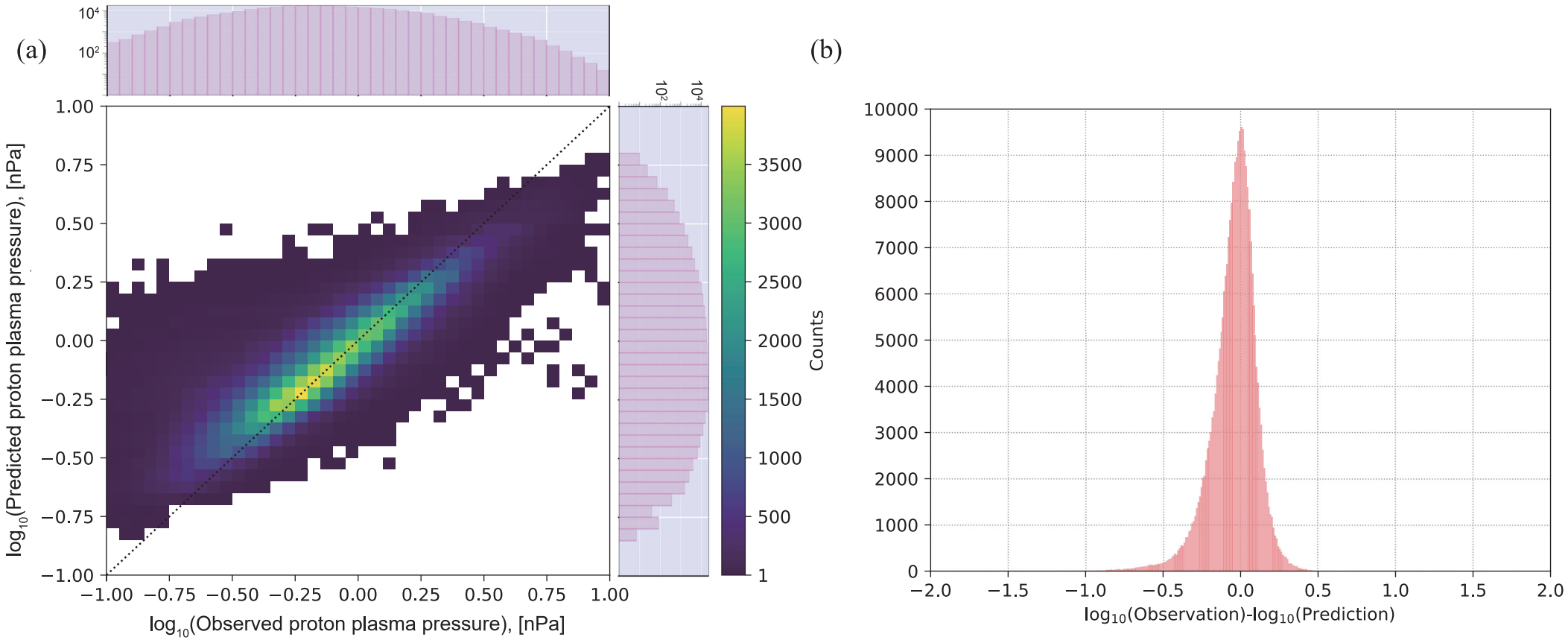

Figure 5. (a) The observed (x-axis) vs. the predicted (y-axis) proton plasma pressure from the training set. The color represents the number of samples in the corresponding bin (10^0.05). The diagonal shows the one-to-one ratio between the observed and predicted pressure.(b)The histogram of the model residuals.

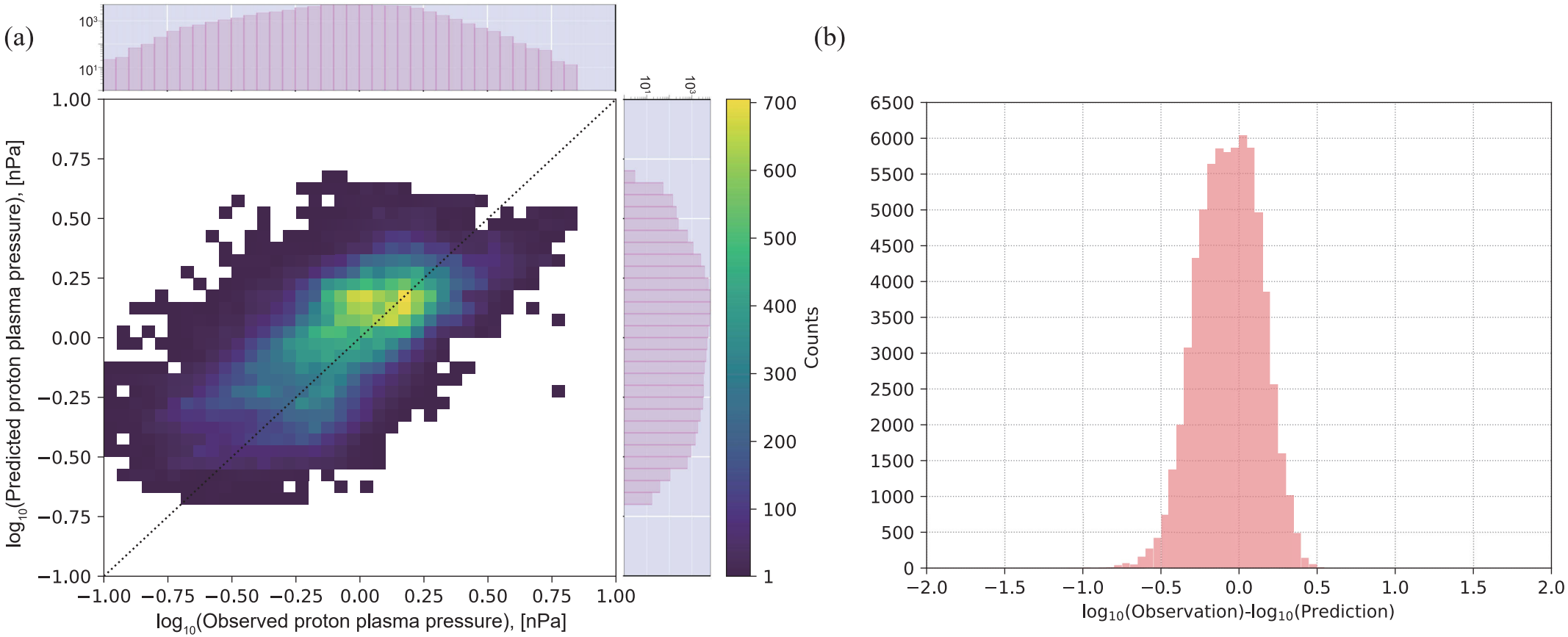

Figure 6. Results from the test set. Same as **Figure5**.

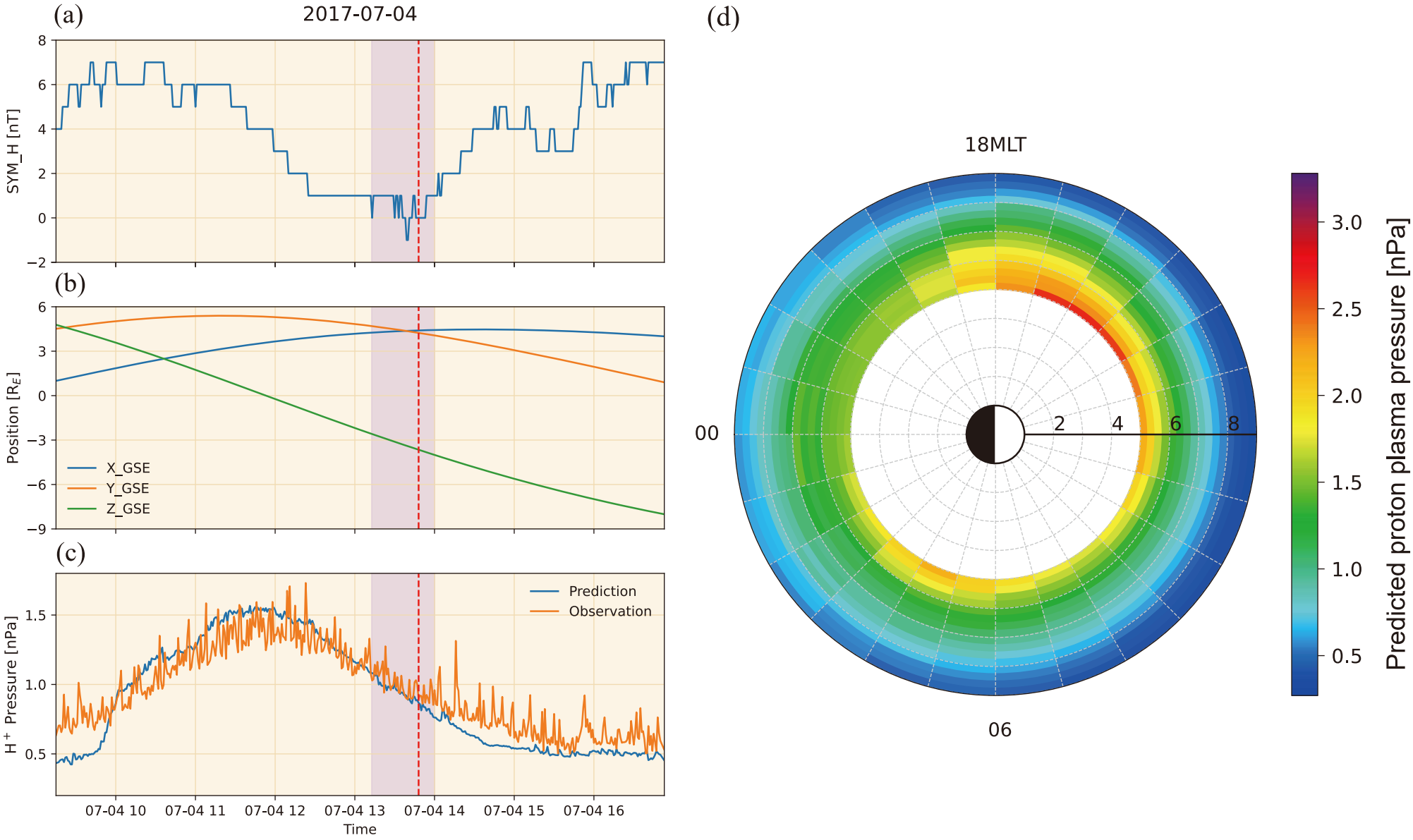

Figure 7. (a)SYM-H index, (b)position, (c)predicted pressure (blue) vs. measured pressure (orange) within the time interval on 2017-07-04 under quiet geomagnetic conditions. (d) Distribution of the predicted H^+ plasma pressure using median values of the parameters over the time period in the purple region as input predictors. MLT is given by the time indicated by the red dotted line.

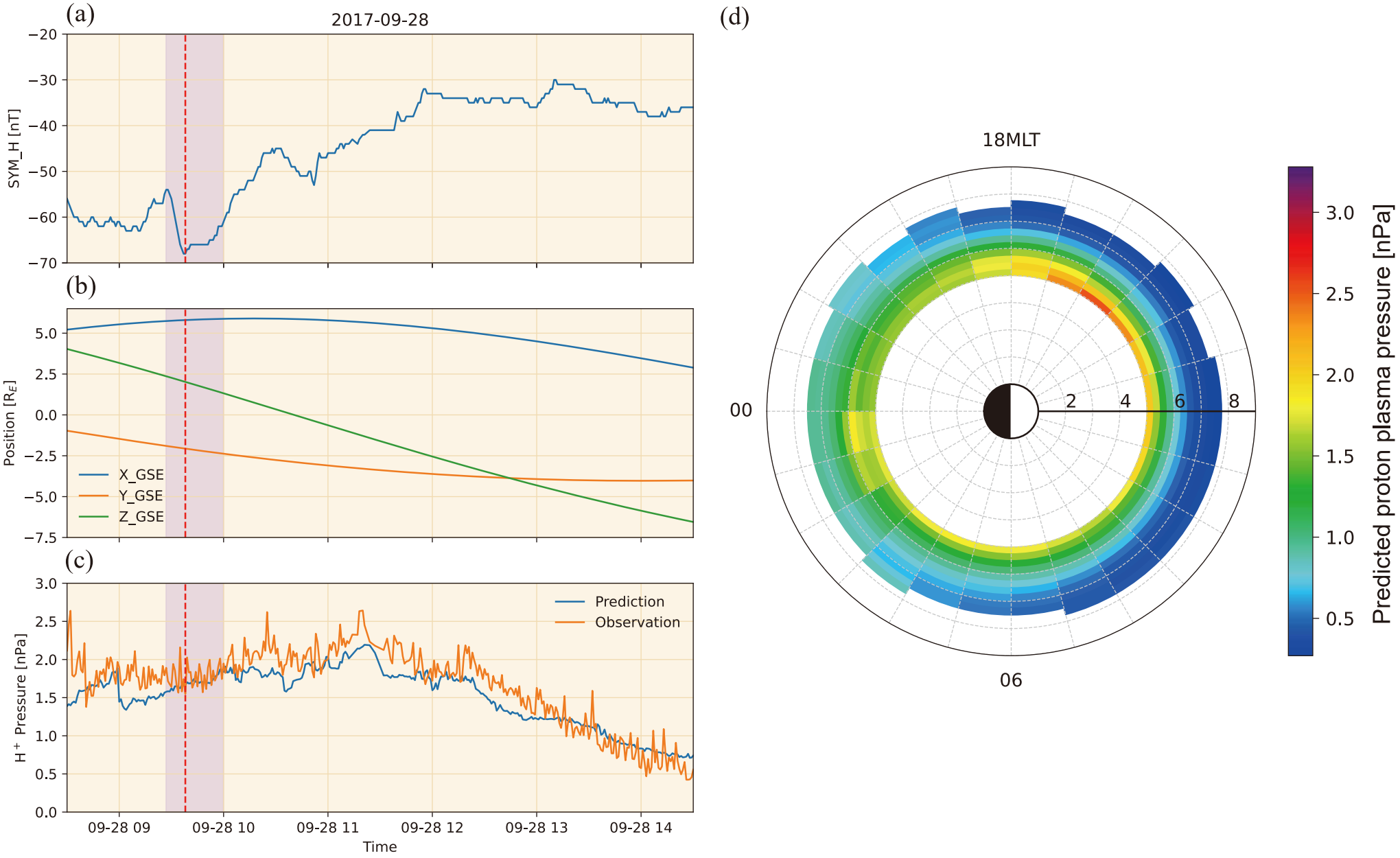

Figure 8. Results within the time interval on 2017-09-28 during the main phase of a geomagnetic storm. Others are the same as **Figure 7**.

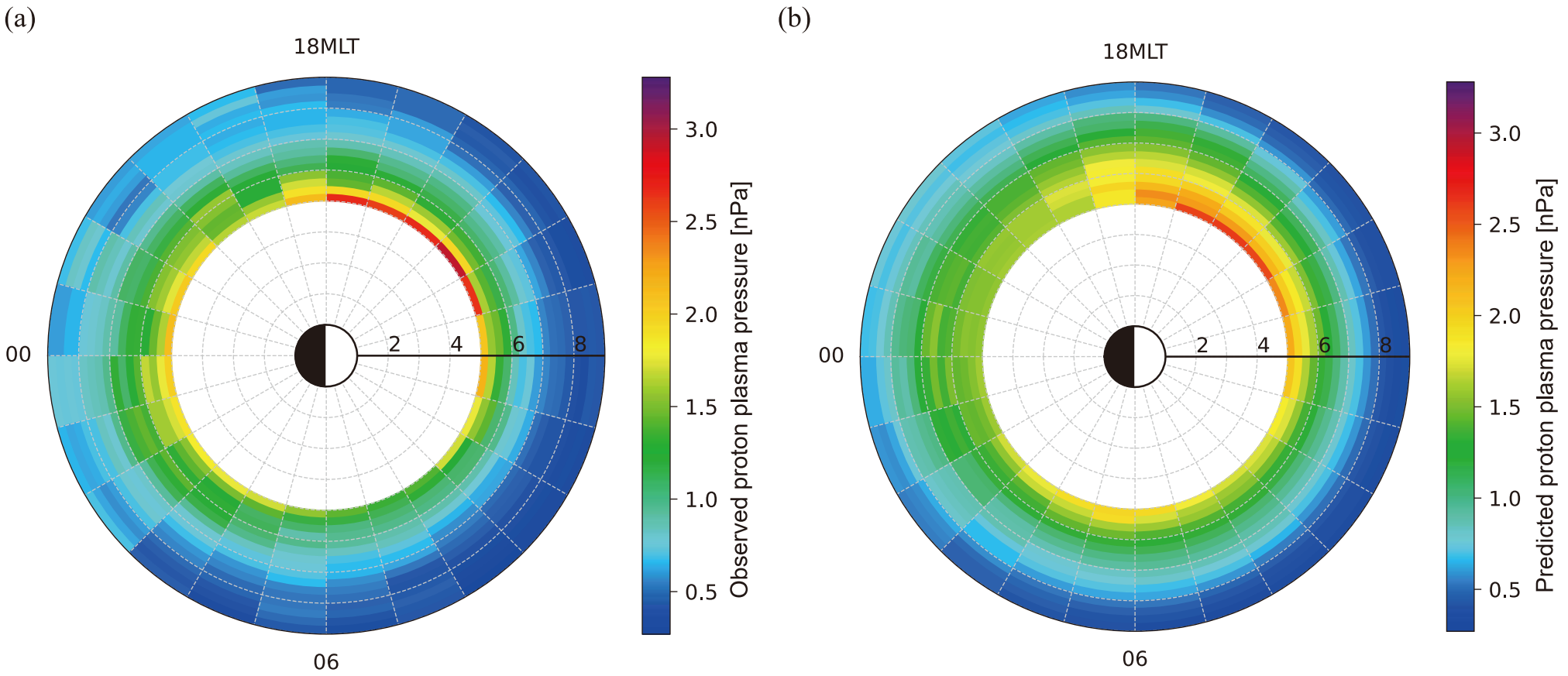

Figure9. (a) Distribution of the observed H^+ plasma pressures by SC4 from February 2001 to February 2018 in the L*-MLT coordinate (Same as **Figure 2a**). (b) Distribution of the predicted H^+ plasma pressures using median values of the parameters over the time range of February 2001 to February 2018 as input predictors in the L*-MLT coordinate.The L* resolution (bin size) is 0.25. MLT resolution (bin size) is 1.

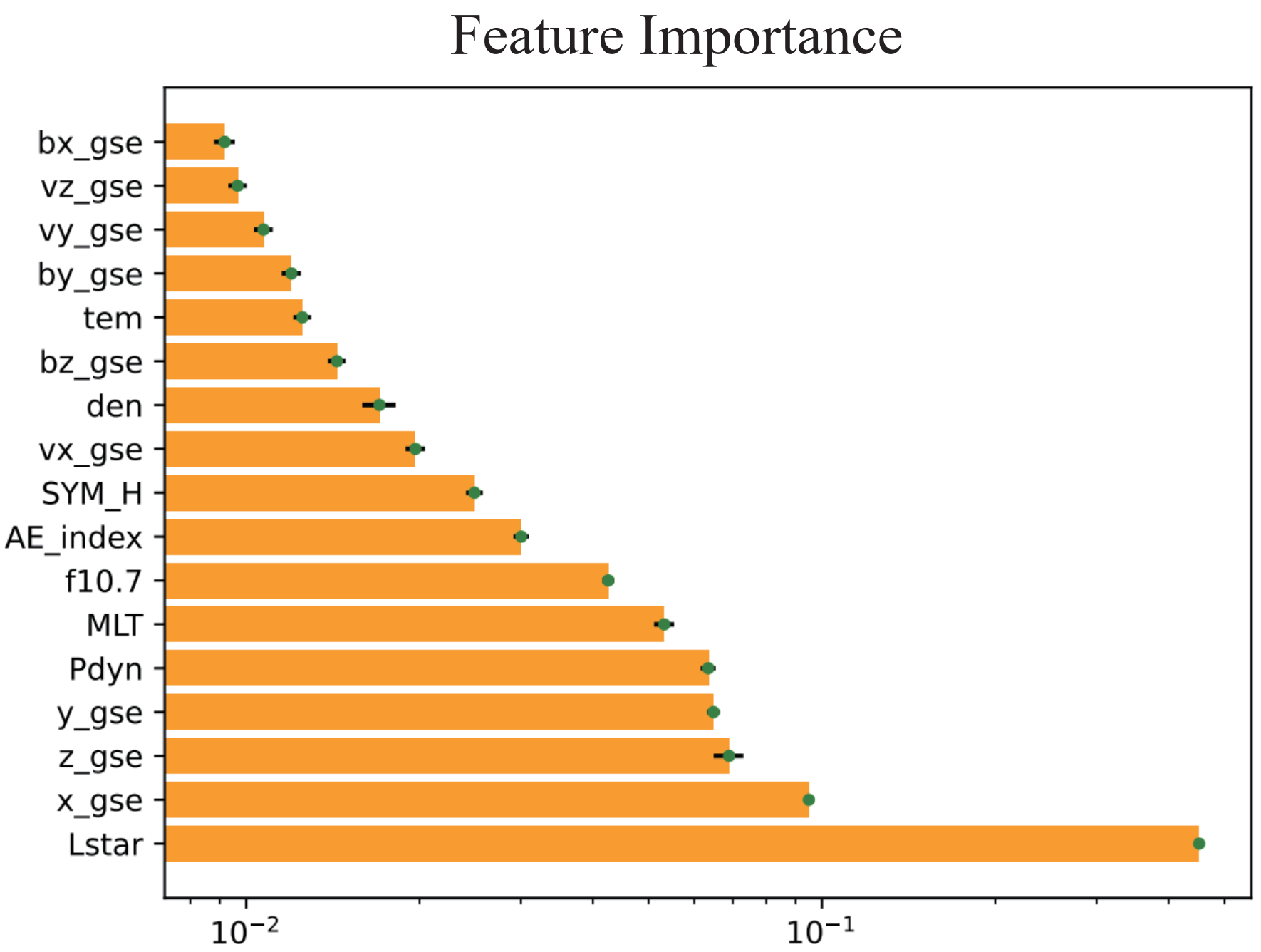

Figure 10. Importance of the different parameters in prediction of H^+ plasma pressure based on training data set. The black horizontal lines represent confidence intervals at 95% confidence level.

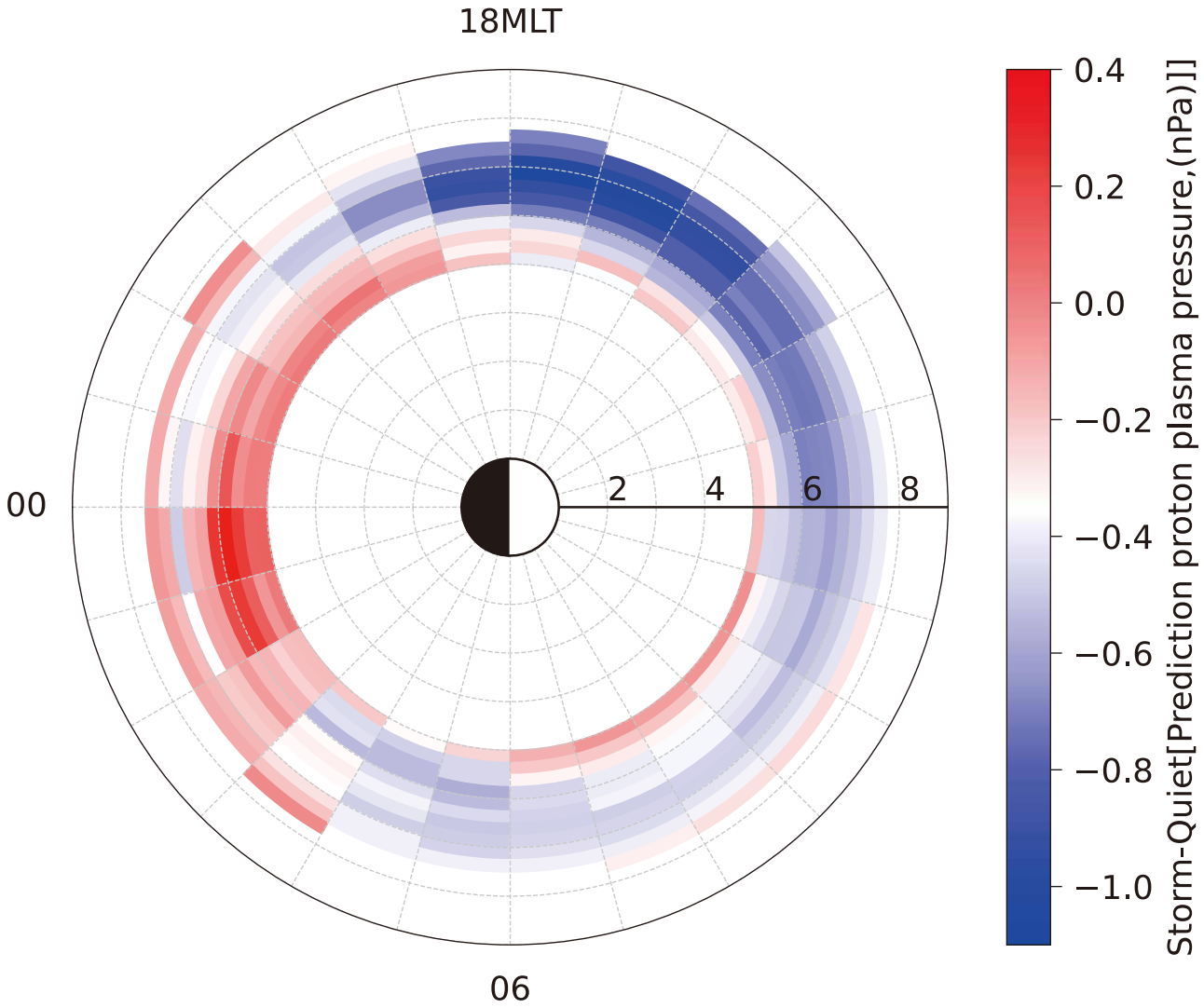

Figure 11. Distribution of the difference between the predicted H^+ plasma pressure under disturbed (shown in **Figure 8d**) and quiet (shown in **Figure 7d**) geomagnetic conditions.

Table 1. Overview of the input features

Feature	Unit	Description
L*		calculated with T89 model
MLT	hr	magnetic local time
x_gse, y_gse, z_gse	RE	position of Cluster in GSE coordinate
Vx/y/z_gse	km/s	components of the solar wind speed in GSE
Bimfx/y/z_gse	nT	components of the interplanetary magnetic field (IMF) in GSE
NpSW	/cm3	solar wind density
Temp	K	solar wind temperature
Pdyn	nPa	solar wind dynamic pressure
AE_index	nT	auroral electrojet index
SYM_H	nT	symmetric H-component index
F10.7	sfu	the solar radio flux at 10.7 cm

Table2. Size and periods of the data subsets after splitting. The time is given in UT.

Subset	Start	End	Number of Points
Train/Validation	2001-02-04 12:31:00	2016-05-16 17:08:00	269184
Test	2016-05-16 17:09:00	2018-02-18 00:02:00	67297

Table 3. Performance of Different Models with Default Input Values

Regressor	Train/Validation Spearman	Test Spearman
ExtraTrees	1.000	0.670
DecisionTree	1.000	0.469
RandomForest	0.996	0.618
LGBM	0.849	0.661
HistGradientBoosting	0.849	0.658
GradientBoosting	0.780	0.664
LinearSVR	0.652	0.634
RidgeRegression	0.648	0.642
LARSRegression	0.645	0.633
AdaBoost	0.628	0.611

Table 4. ExtraTrees Hyperparameters Used for Model Setup

Name	Search range	Value
n_estimators	10-350	140
max_depth	1-30	13
min_samples_leaf	1-50	35

Note. Other parameters use default values.

Table 5. Performance of the Model for Trained/Validation and Test Data Sets

Metrics	Train	Validation	Test
Spearman	0.89	0.71	0.68
MSE	0.12	0.39	0.31
MAE	0.21	0.39	0.40
r^2	0.79	0.13	0.26

Appendix

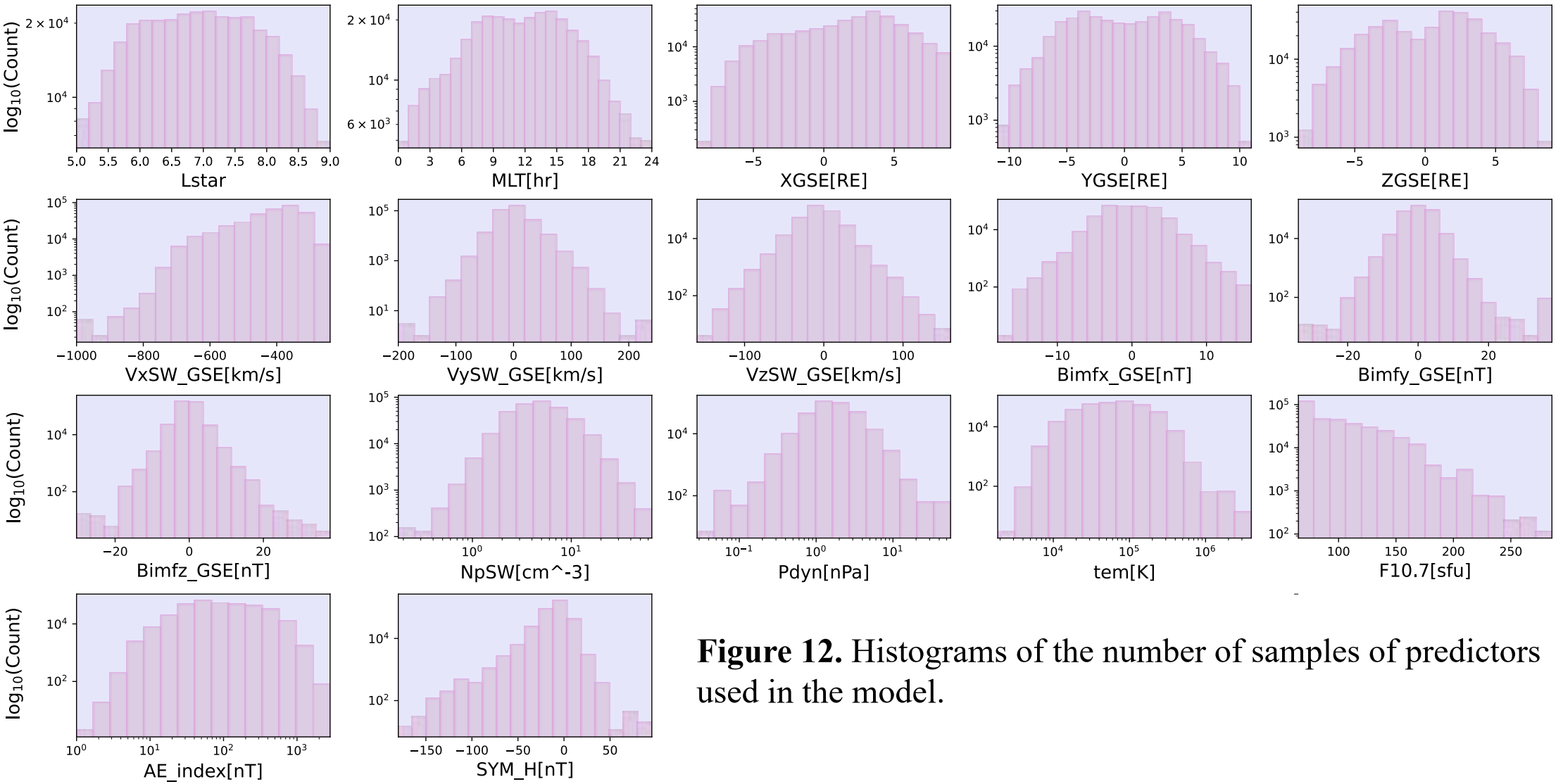

Figure 12. Histograms of the number of samples of predictors used in the model.

Table 6. Comparison of median values of predictors during different time periods

Parameter	Median of quiet time	Median of storm time	Median of all points
SYM_H index (nT)	1	-65.5	-8
Bx_gse (nT)	0.215	-2.49	-0.23
By_gse (nT)	-2.675	0.83	0.11
Bz_gse (nT)	-0.54	3.41	-0.01
Vx_gse (km/s)	-342.8	-687.05	-403.2
Vy_gse (km/s)	-6.6	28.5	-1.2
Vz_gse (km/s)	4.75	-28.9	-5.5
Density (/cm^3)	7.81	3.84	4.7
Temperature (K)	15381.5	263686	63461
Dynamic pressure (nPa)	1.835	3.675	1.64
AE index (nT)	118	581.5	87
F10.7 (sfu)	74.2	91.2	94.6