



# Prediction and Understanding of Soft-proton Contamination in XMM-Newton: A Machine Learning Approach

Elena A. Kronberg<sup>1</sup>, Fabio Gastaldello<sup>2</sup>, Stein Haaland<sup>3,4</sup>, Artem Smirnov<sup>5,6</sup>, Max Berrendorf<sup>7</sup>, Simona Ghizzardi<sup>2</sup>, K. D. Kuntz<sup>8</sup>, Nithin Sivadas<sup>9</sup>, Robert C. Allen<sup>10</sup>, Andrea Tiengo<sup>2,11,12</sup>, Raluca Ilie<sup>13</sup>, Yu Huang<sup>13</sup>, and Lynn Kistler<sup>14</sup>

<sup>1</sup> Department of Earth and Environmental Sciences (Geophysics), University of Munich, Theresienstr. 41, Munich, D-80333, Germany; [elena.kronberg@lmu.de](mailto:elena.kronberg@lmu.de)

<sup>2</sup> Instituto di Astrofisica Spaziale e Fisica Cosmica (INAF-IASF), Milano, via A. Corti 12, I-20133 Milano, Italy

<sup>3</sup> Birkeland Centre for Space Science, University of Bergen, Allégaten 55, NO-5007 Bergen, Norway

<sup>4</sup> Max Planck Institute for Solar System Research, Justus-von-Liebig-Weg 3, Göttingen, Germany

<sup>5</sup> German Research Centre for Geosciences, Albert-Einstein-Straße 42-46, Potsdam, D-14473, Germany

<sup>6</sup> Geophysical Center of the Russian Academy of Sciences, Molodezhnaya St. 3, 119296 Moscow, Russia

<sup>7</sup> Institute of Informatics, University of Munich, Oettingenstraße 67, Munich, D-80538, Germany

<sup>8</sup> Henry A. Rowland Department of Physics & Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

<sup>9</sup> Department of Electrical and Computer Engineering, Boston University, 8 Saint Mary's Street, Boston, MA 02134, USA

<sup>10</sup> Johns Hopkins University Applied Physics Lab, 11100 Johns Hopkins Road, Laurel, MD 20723, USA

<sup>11</sup> Scuola Universitaria Superiore IUSS Pavia, piazza della Vittoria 15, I-27100 Pavia, Italy

<sup>12</sup> INFN, Sezione di Pavia, via A. Bassi 6, I-27100 Pavia, Italy

<sup>13</sup> University of Illinois at Urbana-Champaign, 306 N. Wright Street, 5054 ECEB, Urbana, IL 61801, USA

<sup>14</sup> Space Science Center, University of New Hampshire, Morse Hall Rm 408, Durham, NH 03824, USA

Received 2020 May 28; revised 2020 September 17; accepted 2020 September 23; published 2020 November 6

## Abstract

One of the major and unfortunately unforeseen sources of background for the current generation of X-ray telescopes are few tens to hundreds of keV (soft) protons concentrated by the mirrors. One such telescope is the European Space Agency's (ESA) X-ray Multi-Mirror Mission (XMM-Newton). Its observing time lost due to background contamination is about 40%. This loss of observing time affects all the major broad science goals of this observatory, ranging from cosmology to astrophysics of neutron stars and black holes. The soft-proton background could dramatically impact future large X-ray missions such as the ESA planned Athena mission (<http://www.the-athena-x-ray-observatory.eu/>). Physical processes that trigger this background are still poorly understood. We use a machine learning (ML) approach to delineate related important parameters and to develop a model to predict the background contamination using 12 yr of XMM-Newton observations. As predictors we use the location of the satellite and solar and geomagnetic activity parameters. We revealed that the contamination is most strongly related to the distance in the southern direction,  $Z$  (XMM-Newton observations were in the southern hemisphere), the solar wind radial velocity, and the location on the magnetospheric magnetic field lines. We derived simple empirical models for the first two individual predictors and an ML model that utilizes an ensemble of the predictors (Extra-Trees Regressor) and gives better performance. Based on our analysis, future missions should minimize observations during times associated with high solar wind speed and avoid closed magnetic field lines, especially at the dusk flank region in the southern hemisphere.

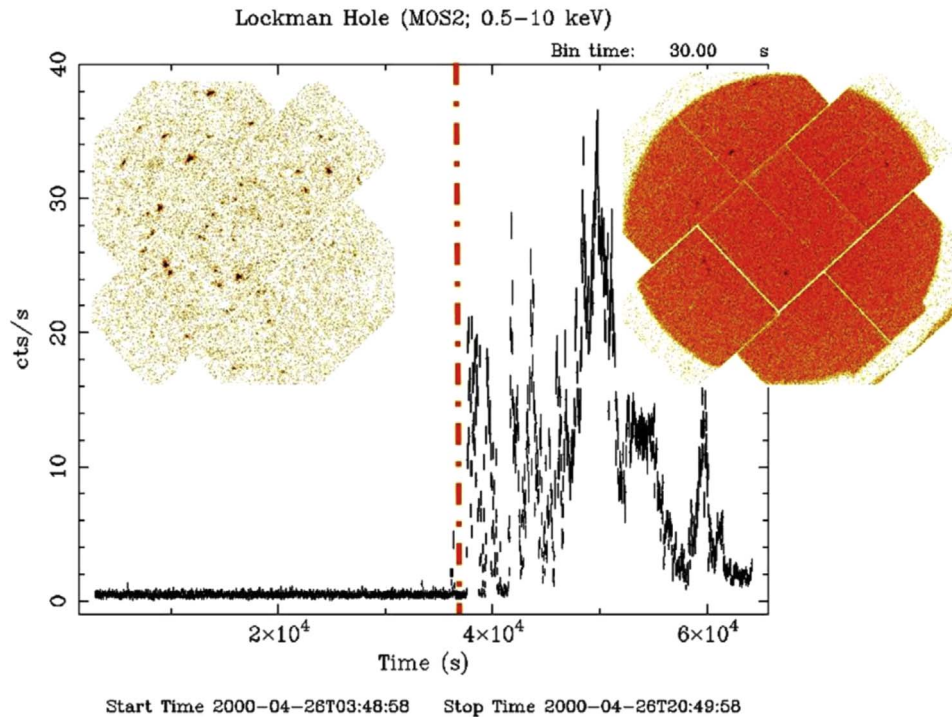
*Unified Astronomy Thesaurus concepts:* X-ray telescopes (1825); X-ray detectors (1815); X-ray observatories (1819); Space plasmas (1544); Astronomy data modeling (1859); Astronomy data analysis (1858)

## 1. Introduction

X-ray telescopes are built to focus X-ray photons toward the detectors in the focal plane by a double low-angle scattering (grazing incidence) from concentric mirror shells. For the past two decades, with the advent of the modern X-ray observatories in orbit such as Chandra (see, e.g., Weisskopf et al. 2002) and XMM-Newton (Jansen et al. 2001), it has been recognized that protons of energies in the range of tens of keV up to a few MeV, hereafter referred to as soft protons (SP), can scatter at low angles through the mirror shells and reach the focal plane (see, e.g., Fioretti et al. 2016 and references therein). These protons, populating the interplanetary space and Earth's magnetosphere, can damage CCD detectors by delivering a nonionizing dose, leading to a loss of spectral resolution. Their signal is indistinguishable from X-ray photons. Therefore, it cannot be rejected, and it produces an enhanced background.

This phenomenon was discovered after the damaging of the Chandra/Advanced CCD Imaging Spectrometer (ACIS)

front-illuminated (FI) CCDs during its first passage through the radiation belt (Prigozhin et al. 2000b). Analysis of data of the calibration source showed that all the FI CCD chips had suffered some damage, causing a significant increase in the charge transfer inefficiency (CTI). CTI is caused by defects in the silicon lattice that can be created by the interaction with charged particles. These defects, or “traps,” capture charges during their transfer to the read-out electronics and release them at later times. Their effects on the detector performance are position-dependent changes in the energy scale, loss of spectral resolution, and loss of quantum efficiency (O'Dell et al. 2000; Prigozhin et al. 2000a). Therefore, after less than 2 months of operation ACIS has been protected during radiation belt passages, by moving the detector out of the telescope focus (O'Dell et al. 2003). The same procedure takes place during periods of enhanced particle flux, triggered either by the onboard radiation monitor or by ground operations monitoring of various space weather probes (Grant et al. 2012).



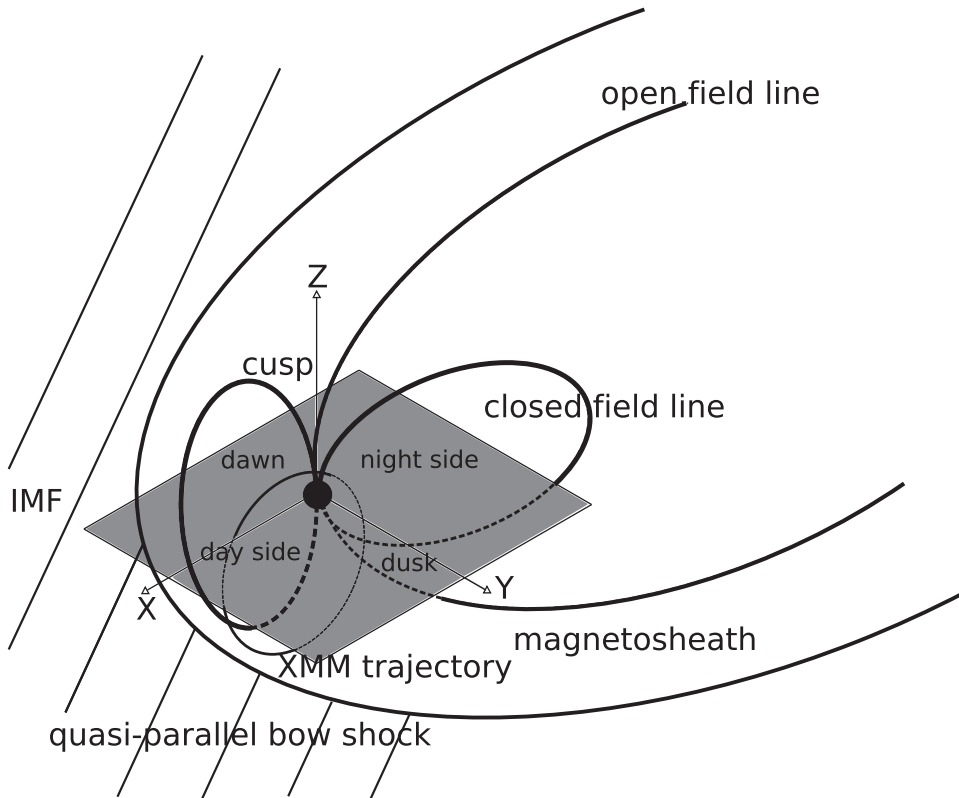
**Figure 1.** Example of XMM-Newton observation partly affected by SPs. The flares are clearly visible in the second part of this light curve taken from MOS2, one of the detectors on board XMM. Their effect on the exposure quality can be evaluated comparing the image extracted from the first half (left) and second half (right) of the observation. Adapted from Lotti et al. (2018).

XMM-Newton was launched into an orbit similar to Chandra, only with apogee in the southern hemisphere. To avoid radiation belts, the detectors of XMM-Newton are kept closed with a  $\sim 1$  mm thick aluminum shield below altitudes of about 40,000 km. XMM-Newton’s highly eccentric elliptical orbit, with an apogee of about 115,000 km and a perigee of about 6000 km from Earth, traverses the full range of magnetospheric environments, from the inner magnetosphere to the solar wind (SW) when the satellite is outside the bow shock. Along its orbit the satellite encounters enhanced intensities of SPs. These episodes are hereafter referred to as “SP flares.” They occur on extremely variable timescales, ranging from hundreds of seconds to several hours. The peak count rate can be more than three orders of magnitude higher than the quiescent one (De Luca & Molendi 2004). The extreme time variability is the fingerprint of this background component, the so-called SP component, which should not be confused with solar flares or solar energetic particles. A light curve can immediately show the time intervals affected by a high background count rate. Such intervals are usually not suitable for scientific analysis unless the X-ray source to be studied is extremely bright (see Figure 1). They have to be rejected, discarding all of the time intervals having a count rate above a selected threshold.

A preliminary analysis of the distribution of flares as a function of orbital position, distance from Earth, and orbital phase with respect to the Sun has been done by Kuntz & Snowden (2008). The part of the orbit that seems the most susceptible to SP flare is in the inner part of the magnetosphere (near perigee), whereas the greatest flare-free time occurs when the spacecraft is farthest from Earth, either outside the bow shock or deep within the magnetotail (Ghizzardi et al. 2017). A

development of that work based on XMM-Newton measurements from 2000 to 2010 for a total of 51 Ms of data concluded that the highest percentage of proton flares occurred when the spacecraft is on closed magnetic field lines (Walsh et al. 2014); see sketch in Figure 2. According to this study, the SPs affect  $\sim 55\%$  of measurements, and that can be as high as 66% of measurements when XMM-Newton is located in low-latitude magnetospheric regions on closed magnetic field lines. Other studies (e.g., Salvetti et al. 2017) report a mean contamination rate of  $\sim 35\%$ . A recent analysis based on about 100 Ms of data measured between 2000 and 2012 confirmed the general trend of a decreasing intensity with distance from Earth (shown by the mean count rate of the SP component). It also showed that the dayside magnetosphere with closed field lines is more contaminated by SP flares than regions on the night side on open field lines (Ghizzardi et al. 2017).

The performance of future X-ray focusing telescopes orbiting in the interplanetary space will suffer from SP-induced background events. Of particular concern is European Space Agency’s (ESA) next large-class mission ATHENA (Nandra et al. 2013), given that its large effective area ( $1.4 \text{ m}^2$  at 1 keV) makes the minimization of SP contamination a key challenge for the fulfillment of ATHENA’s science objectives, as is explicitly recognized in the background requirements of the mission. A possible shielding solution is placing an array of magnets (a magnetic diverter) between the optics and the focal plane, able to deflect charged particles away from the instruments’ field of view (e.g., Fioretti et al. 2018; Lotti et al. 2018). The initial choice of an L2 orbit is also being reconsidered owing to the far superior knowledge of the various proton components in L1 (Fioretti et al. 2018; Laurenza et al. 2019). The first and second Sun–Earth Lagrange points



**Figure 2.** Sketch of the terrestrial magnetosphere; oblique lines in front of the magnetosphere represent the interplanetary magnetic field (IMF), and X, Y, and Z denote directions of the Geocentric Solar Ecliptic (GSE) coordinate system. XMM-Newton apogee is found at  $\sim 18R_E$ , where  $R_E$  is Earth’s radius. In the time period considered here, the XMM-Newton orbit has changed from highly elliptical to more circular and then back to highly elliptical.

(L1 and L2) are locations where the gravitational forces of the Sun and Earth cancel. Both L1 and L2 are located along the Sun–Earth line, with L1 being  $1.5 \times 10^6$  km sunward of Earth, while L2 is located at the same distance behind Earth.

In this paper we delineate which of the geometric, solar, SW, and geomagnetic parameters mostly control strong contamination in the XMM-Newton telescope using a machine learning (ML) approach. The eventual aim is to define the cause of the contamination. The ML approach has been successfully used to predict plasma environments in the terrestrial magnetosphere, such as electron density in the plasmasphere (Zhelavskaya et al. 2017) and the inner magnetosphere (Chu et al. 2017) and the electron intensity in the radiation belts (Smirnov et al. 2020) and in the SW (Roberts et al. 2020). The advantage of this approach is that it allows complex nonlinear relationships to be analyzed in large data sets (Geron 2019). Our task is to predict target numeric values, namely, the count rate of the SP contamination, given a set of features, such as location of the satellite and solar, SW, and geomagnetic parameters, called predictors. We treat this problem as a regression (see, e.g., Camporeale 2019 for details). To train the algorithm, one feeds it with many examples of events that include both their predictors and their desired solutions (count rates of the contamination in our case). Such an ML approach is called supervised learning; the training set given to the algorithm includes the desired solution. Some of the most important supervised learning algorithms are linear regression, support vector machines, decision trees and random forests, neural networks, gradient descent, and gradient boosting.

To predict the contamination, we first explored the relation with the single parameters to help select the best predictors for an ML model. With this choice we test a row of supervised ML algorithms and eventually derive a model that utilizes an ensemble of predictors based on the Extra-Trees Regressor algorithm. Using this ML algorithm, we evaluated the importance of nonlinear relationships.

The ML approach may help in searching for similar patterns between the XMM-Newton contamination and SP intensities measured by Cluster, thus constraining the source of SPs. ESA’s mission Cluster, which is a suite of four satellites (Escoubet et al. 1997), orbits Earth on polar trajectories similar to XMM. However, there are no physical conjunctions between these two satellites that would allow direct insights on what exactly produces contamination. Therefore, in the future one approach will be the identification of possible magnetic field conjunctions (observations at similar magnetic field topologies) and comparing observations from both missions. Another approach will be to delineate which geometric, solar, SW, and geomagnetic parameters are most related to dynamics of SPs at different energies observed by Cluster to compare with those parameters associated with the XMM-Newton contamination. In the future, we will derive an ML predicting model for the SPs measured by Cluster and apply it to XMM-Newton trajectories to disentangle at which energies SPs are best correlated with the contamination. By this we will determine the energy of SPs that contaminates the detector the most.

This work has been inspired by the interdisciplinary collaboration between the astrophysicists and specialists in

magnetospheric physics, supported by the International Space Science Institute in Bern, Switzerland.<sup>15</sup>

## 2. Contamination SP Count Rates and Their Predictors: Simple Relations

In this section we give details about SP contamination count rates and their predictors. We plot their relations and analyze cross-correlations in order to get better insights into physical processes possibly responsible for the contamination and to have better preselection of the predictors for the ML model.

### 2.1. Contamination Count Rates

The description of the XMM-Newton data set and the analysis performed have been reported in more detail in Marelli et al. (2017) and Salvetti et al. (2017). Here we give a brief and concise summary for the purpose of this paper. The work exploited here has been produced in the framework of AREMBES (ATHENA Radiation Environment Models and X-Ray Background Effects Simulators), which is an ESA project aimed at characterizing the effects of focused and nonfocused particles on ATHENA detectors, both in terms of contributions to their instrumental background and as a source of radiation damage.<sup>16</sup> XMM-Newton is a test bed of the various background components that will be relevant for the ATHENA mission. To this aim we used the XMM-Newton public data set that was available when we started our analysis to produce the most clean data set ever used to characterize the XMM-Newton particle-induced background, taking as input the preliminary results of the FP7 European project EXTraS (Exploring the X-ray Transient and variable Sky;<sup>17</sup> De Luca et al. 2017). The results from Data Release 4 of the 3XMM catalog were required to evaluate the contamination from celestial sources.

The main XMM-Newton instrument is the European Photon Imaging Camera (EPIC), consisting of two metal–oxide–silicon (MOS) detectors (Turner et al. 2001) and a pn camera (Strüder et al. 2001), which operate in the 0.2–12 keV energy range. The EPIC background can be separated into particle, photon, and electronic noise components (see Carter & Read 2007 and Gastaldello et al. 2017 for a detailed description). Aiming to characterize the SP component that is focused by the X-ray telescopes, the key feature exploited is the ability to define in the MOS detectors two detector areas: the in-field-of-view (inFOV) one, exposed to focused X-ray photons and SPs, and the out-field-of-view (outFOV) one, not exposed to sky photons or SPs. The other main component of the particle background, secondary electrons generated by galactic cosmic rays, affects in the same way both the inFOV and outFOV areas of the MOS detectors. The choice of the energy band in the analysis (7–9.4 keV and 11–12 keV) minimizes to a negligible contribution the sky photon component. We focused on MOS2 because we can exploit the full detector area (MOS1 suffered a loss of two of its seven CCDs during the lifetime of the mission).

We can then use the inFOV subtracted by outFOV diagnostic to fully characterize the inFOV excess particle background employing the outFOV region as a calibrator to minimize any contamination. After standard data preparation

and reduction, all the single observations were merged in a final global data set used in this work with 500 s time bins, where the count rate is the difference between the inFOV and outFOV count rate. The work done in the AREMBES project showed two distinct components in the differential distribution of the inFOV–outFOV count rates, one associated with the flares of SPs and the other with a low-intensity component, possibly related to Compton interactions of hard X-ray photons. This fundamental distinction is supported by the comparative analysis of data collected with different filters and a spectral analysis (Salvetti et al. 2017).

We investigate the dynamics of the SP count rates between 0.04 and 200 counts s<sup>−1</sup>. We slightly revise the lowest threshold of 0.1 counts s<sup>−1</sup> used in Ghizzardi et al. (2017). A threshold of 0.1 was chosen in Ghizzardi et al. (2017) to be in a regime totally dominated by the SP contribution. However, the regime between 0.04 and 0.10 still provides a significant contribution with respect to the other major component of the XMM-Newton background, which is the galactic-cosmic-ray-induced background (which ranges in the same units from 0.1 to 0.4; see left panel of Figure 4 in Salvetti et al. 2017). We select the observations with radial distance above 6R<sub>E</sub>. From 2001 January 2 to 2012 August 30, 707,330 minutes of data matched these criteria. We also applied the base-10 logarithm to the SP count rates because the data variation is in the range of several orders of magnitude. The distribution of the number of samples for the predictors and the count rates (on the vertical axis) with a given value range (on the horizontal axis) is shown in Figure A1 in the Appendix.

### 2.2. Predictors Related to Location in Space

Each count rate was associated with location in the Geocentric Solar Ecliptic (GSE) coordinate system, represented by parameters  $X$ ,  $Y$ ,  $Z$  (see sketch in Figure 2), and the radial distance from Earth, parameter  $rdist$ . Throughout the paper distances are given in  $R_E$  units. The distribution of the SP count rates in the GSE system is shown in Figure 3. The figure shows that the day side (positive XGSE) is more affected by the contamination. A duskward asymmetry is observed, with stronger contamination toward flanks on the dusk side and higher count rates at approximately YGSE = 8  $R_E$  and XGSE 6–12  $R_E$  (see sketch in Figure 2 that indicates location of the dusk/dawn and day/night sides). Figure 3 illustrates a decrease of SP contamination at larger distances from Earth in the  $Z$  direction.

In Figure 4 we plot count rates versus individual predictors. One can see that the logarithm of the SP count rates almost linearly decreases with  $Z$ ; see Figure 4(a). This dependence is the strongest compared to the other parameters, considering the span of the count rate values. The linear regression derived from this dependence (shown by the red line in Figure 4(a)) is

$$\log_{10}(\text{SP Count rates}) = 0.328 + 0.0725 \cdot Z.$$

This relation indicates an exponential dependence of the SP count rates on  $Z$ . For this linear regression the Pearson correlation,  $r$ , is 0.99, the probability value,  $p$ , is  $4 \times 10^{-11}$ , and the standard error of the estimated gradient is  $3 \times 10^{-3}$ .

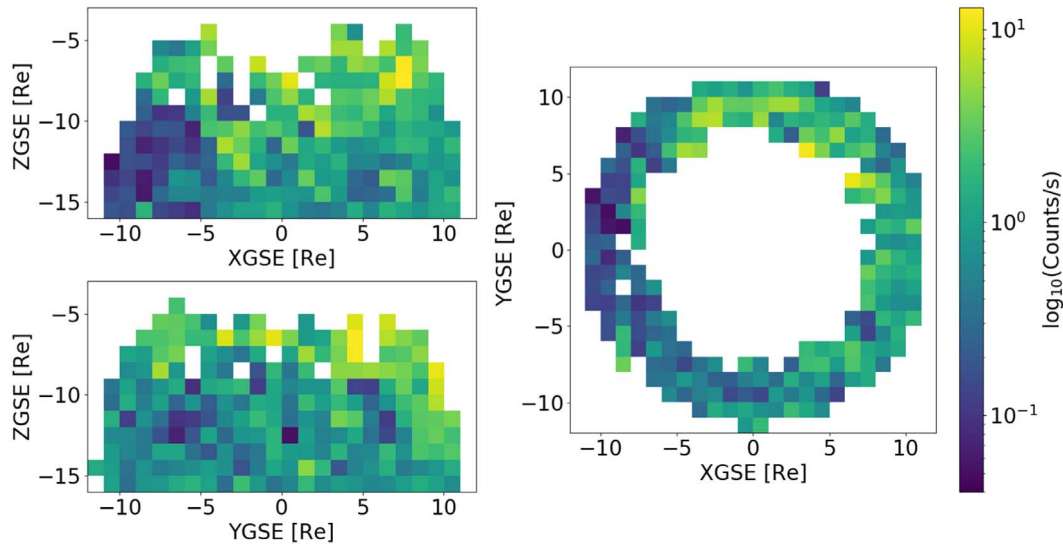
The change of the logarithm of count rates with  $Y$  is nonlinear and significantly weaker than those on  $Z$ ; see Figure 4(b). The duskward asymmetry expected from Figure 3 is clearly observed. A slightly less strong, nonlinear dependence of logarithm of count rates is seen with respect to  $X$

<sup>15</sup> <https://www.issibern.ch/teams/softprotonmagxray/>

<sup>16</sup> <http://space-env.esa.int/index.php/news-reader/items/AREMBES.html>

<sup>17</sup> <http://www.extras-fp7.eu/>





**Figure 3.** Distribution of the SP count rates in the range between 0.04 and 200 counts  $\text{s}^{-1}$  in the GSE coordinate system. The number of SP count rates per bin is larger than 2.

(Figure 4(c)), with a higher level of contamination along the day side, as also seen from Figure 3.

Previous studies (e.g., Walsh et al. 2014) have demonstrated that the XMM-Newton SP contamination count rates depend on the type of the connection of the magnetic field line to Earth. Therefore, we have added a parameter called *Foot Type*: closed magnetic field lines with both ends at Earth (*Foot Type* = 2), open magnetic field lines with one end at Earth and the other end connected to the interplanetary magnetic field (IMF) (*Foot Type* = 1), and magnetic field lines not connected to Earth, namely, IMF (*Foot Type* = 0); see sketch in Figure 2. The parameter *Foot Type* also describes the location of the contamination with respect to Earth’s magnetosphere. This parameter was calculated using the Tsyganenko 96 model (Tsyganenko 1995). There are later versions of the Tsyganenko model that may be more appropriate in periods of high SW dynamic pressure. However, for practical reasons we use only one model. In Figure 4(d) we can see that there are significantly higher count rates on the closed field lines (*Foot Type* = 2) than either on open field lines (*Foot Type* = 1) or on IMF field lines (*Foot Type* = 0). Additionally, the count rates in the IMF (*Foot Type* = 0) are also significantly higher than on the open field lines (*Foot Type* = 1).

### 2.3. Predictors Related to the Solar, Solar Wind, and Geomagnetic Activity

The XMM-Newton count rates were combined with simultaneous observations of the solar, SW, and geomagnetic parameters taken from the OMNI database;<sup>18</sup> see also King & Papitashvili (2005). The SW observations are taken from the OMNI data set. They are propagated to Earth’s bow shock. The SW is characterized by the proton density,  $Np_{\text{SW}}$  in  $\text{cm}^{-3}$  (see Figure 4(g)); components of the speed in the GSE coordinates,  $V_{x\text{SW\_GSE}}$ ,  $V_{y\text{SW\_GSE}}$ , and  $V_{z\text{SW\_GSE}}$  in  $\text{km s}^{-1}$  (see Figure 4(e) for the former component); the temperature,  $Temp$ , in K (see Figure 4(f)); the dynamic pressure,  $P_{\text{dyn}}$ , in nPa, which is calculated as  $Np_{\text{SW}} V_{\text{SW}}^2 \times 1.67 \times 10^6$  (see Figure 4(h)); components of the IMF in the GSE coordinates,

$B_{\text{imfxGSE}}$ ,  $B_{\text{imfyGSE}}$ , and  $B_{\text{imfzGSE}}$ , in nT (see Figure 4(j)); and clock angle (CA) calculated as  $\arctan(B_{\text{imfyGSE}}/B_{\text{imfzGSE}})$ . To consider the influence of solar irradiation, we included the F10.7 index, which measures the radio flux at 10.7 cm (2.8 GHz; Tapping 2013). This parameter correlates well with the sunspot number and other indicators of solar and UV solar irradiance and can be measured reliably under any terrestrial weather condition (unlike many other solar indices). It is denoted by *F107* and measured in solar flux units (sfu; see Figure 4(i)). The parameters of geomagnetic activity such as the auroral electrojet (AE) index, denoted as  $AE_{\text{index}}$ , in nT, characterizing the magnetic field disturbance in the auroral region of the northern hemisphere, and SYM-H index, denoted as *SYM-H* and measured in nT, characterizing the disturbance of the geomagnetic field at the equatorial regions, are considered (Nose et al. 2017).

In Figure 4 we plot parameters for the most prominent relations with the SP count rates. We acknowledge that the SW and geomagnetic properties are correlated with one another. Thus, we try to determine here which dominates.

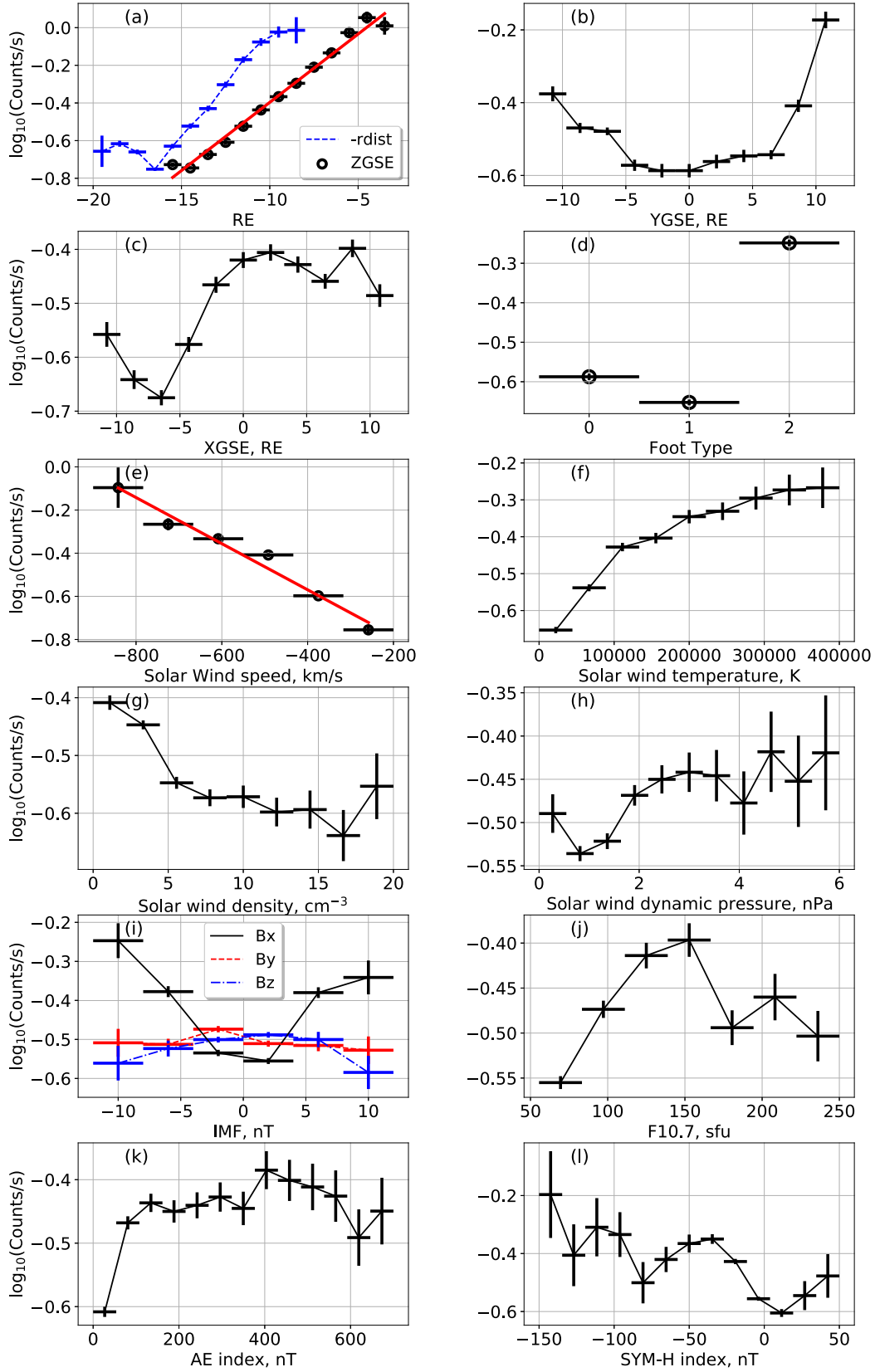
The logarithm of count rates increases almost linearly with absolute value of the SW speed; see Figure 4(e). The linear regression derived from this dependence (shown by the red line in Figure 4(e)) is

$$\log_{10}(\text{SP Count rates}) = -0.997 - 10^{-3} \cdot V_x.$$

This relation indicates an exponential dependence of the SP count rates on  $V_x$ . The  $r$  is  $-0.99$ , the  $p$  value is  $2.5 \times 10^{-4}$ , and the standard error of the estimated gradient is  $8.7 \times 10^{-5}$ .

The count rates clearly increase with the SW temperature; see Figure 4(f). The count rates nonlinearly decrease with the SW density; see Figure 4(g). The SP count rate relation with the SW pressure is nonlinear; see Figure 4(h). For SW pressures higher than 6 nPa, the confidence intervals cover the entire value range of count rates. These values are discarded because they are statistically insignificant. The relation of count rates with the SW pressure is less important than the one with the SW speed because the count rates anticorrelate with the SW density; see Figure 4(g). SW speed is often anti-correlated with

<sup>18</sup> <https://omniweb.sci.gsfc.nasa.gov>



**Figure 4.** Relations of mean XMM-Newton count rates and (a–c) ZGSE together with linear regression shown by the red line and negative radial distance, YGSE and XGSE, respectively; (d) foot Type; (e) the SW radial velocity and its linear regression shown by the red line; (f–h) SW temperature, density, and dynamic pressure, respectively; (i) IMF components in GSE; (j) F10.7 parameter; (k) AE index; and (l) SYM-H index. Vertical lines represent standard in statistics confidence intervals at 95% confidence level. Horizontal lines represent the half width of the bin for which the corresponding values were calculated. The data points are connected by thin lines to guide the eye.

SW density (Richardson 2018), therefore reducing the significance of the SP count rates relative to the SW pressure.

Of the IMF components, the  $B_x$  component shows the strongest relation with the logarithm of count rates; see Figure 4(i). The XMM-Newton count rates increase with the absolute value of the IMF  $B_x$  component. The IMF  $B_y$  component does not show significant change. It is rather unexpected to observe a significant decrease of the SP count rates for absolute values of IMF  $B_z > 8$  nT. The strong values of IMF  $|B_z| > 8$  nT are likely associated with geoeffective interplanetary phenomena such as coronal mass ejections (CMEs) and corotating interaction regions (CIRs) (Gonzalez et al. 1999; McPherron & Weygand 2006; Li et al. 2018), and intuitively one would expect an increase in the count rates. This will be discussed in Section 4.

The logarithm of count rates first increases linearly with the F10.7 index up to 150 sfu and then significantly drops at higher values (see Figure 4(j)), indicating a nontrivial influence of this parameter on level of contamination.

The dependence of the contamination on the substorm activity, indicated by AE index, is weaker than for the SW speed; see Figure 4(k). The SP count rates significantly grow with AE index at least up to 100 nT. In general, AE index, namely, strong magnetic field disturbance in the northern high-latitude region, does not show a significant relation with count rates at values  $>100$  nT.

The dependence of the count rates on the SYM-H index is also nontrivial. The SP count rates increase for decreasing values of the SYM-H index from 0 up to approximately  $-50$  nT; see Figure 4(l). At lower SYM-H index values dependence becomes nonlinear with large error bars. An increasingly negative SYM-H index means that the ring current is stronger at equatorial latitudes.

#### 2.4. Cross-correlations between Contamination Counts and Predictors

In Figure 5 we show the correlation coefficient,  $r$ , between parameters possibly related to the level of SP contamination. The values of Pearson correlation vary between  $-1$  and  $1$ , with values close to  $-1/1$  meaning perfect linear anticorrelation/correlation and values close to  $0$  meaning no linear correlation. We dropped the  $V_ySW\_GSE$  and  $V_zSW\_GSE$  components from this plot for the sake of better presentation, as they show very low correlation with SP counts and small influence on reproducing the counts in the model. These velocity components are small compared to the  $V_xSW\_GSE$ . We also checked correlation and influence of the total SW speed on the reproducibility of the SP count rates; however, it shows a very similar behavior to  $V_xSW\_GSE$ . In order to avoid redundant parameters, we do not include this variable. Additionally, we dropped  $BimfyGSE$  from Figure 5 owing to the low Pearson correlation and no obvious relationship with SP counts in Figure 4. The correlations help us to exclude parameters that are strongly correlated with one another that can overload the model. However, one should be also careful with the interpretation of the Pearson correlation coefficient, as this indicates only linear relationships.

The results of the cross-correlation analysis follow the strongest relations with count rate, such as variation with ZGSE direction ( $r = 0.35$ ), radial distance ( $r = 0.32$ ), Foot Type ( $r = 0.23$ ),  $V_xSW\_GSE$  ( $r = 0.21$ ), SW temperature ( $r = 0.17$ ), and SYM-H index ( $r = 0.12$ ). Here we chose predictors with

correlation larger than  $0.1$ . However, one can also note well-defined nonlinear relations of count rates with  $X$ ,  $Y$ , and  $BimfxGSE$  in Figure 4 that got low scores in Pearson correlation.

On the basis of correlations and dependencies in Figure 4 we select the following predictors for the ML model:  $X$ ,  $Y$ ,  $Z$ ,  $r_{dist}$ ,  $Foot\ Type$ ,  $V_xSW\_GSE$ ,  $P_{dyn}$ ,  $BimfxGSE$ ,  $F107$ ,  $AE\_index$ , and  $SYM-H$ . The  $BimfyGSE$  and  $BimfzGSE$  are dropped because they do not show much variation with the counts. The  $r_{dist}/NpSW/Temp$  are dropped because they correlate strongly with  $Z/P_{dyn}/V_xSW\_GSE$  and do not significantly improve the model.

### 3. ML Model for SP Contamination

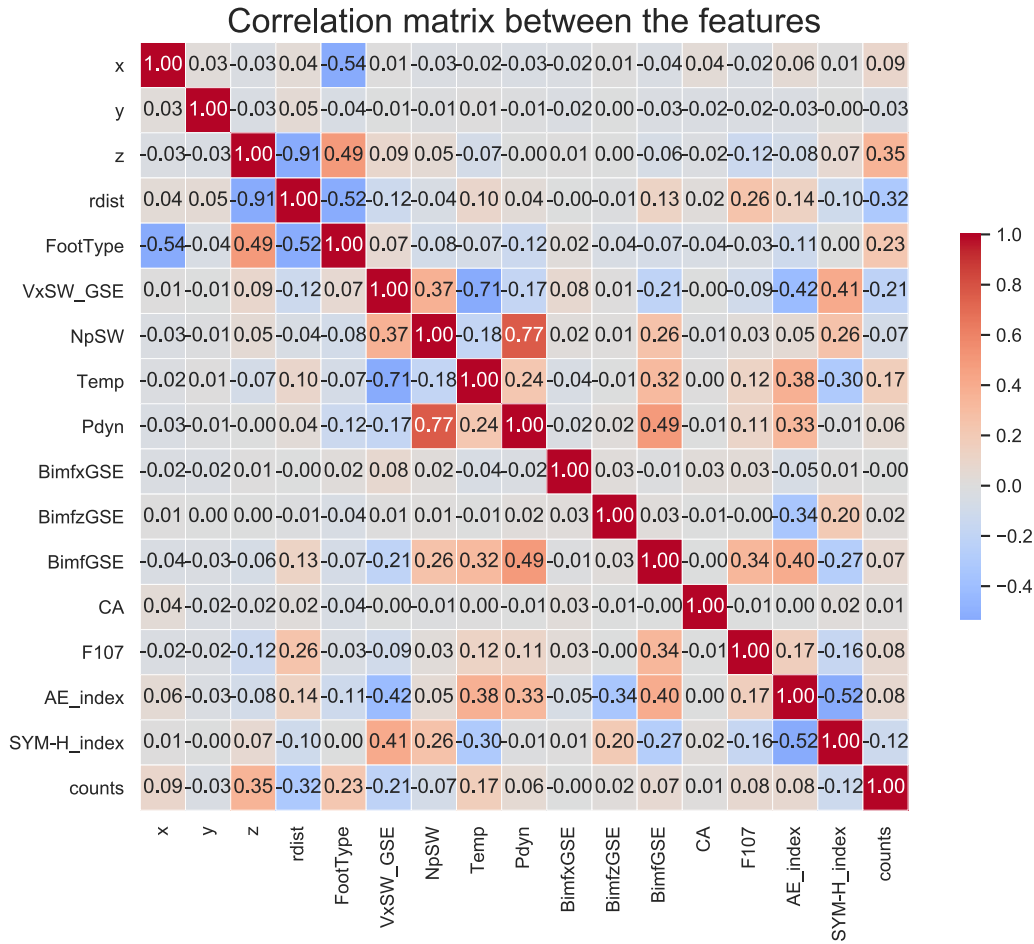
The relation between the SP count rates and the row of different predictors listed above is complex; see Figure 4. It is, therefore, often a group of predictors or their ensemble that gives better predictions than the best individual predictor (Geron 2019).

From supervised ML regressions we have tried Stochastic Gradient Descent Regressor (SGDRegressor), Gradient Boosting for Regression (GradientBoostingRegressor), Random Forest Regressor (RandomForestRegressor), Extra-Trees Regressor (ExtraTreesRegressor), and Multi-layer Perceptron Regressor (MLPRegressor) methods implemented in Scikit-Learn (Pedregosa et al. 2011). These methods show comparable or slightly worse performance; see Table 1. To evaluate performance, we use Spearman correlation,  $\rho$ , between results of the model on training/test data sets and observations that are listed as Train Spearman/Test Spearman in Table 1, respectively. The values of Spearman correlation vary between  $-1$  and  $1$ , with values close to  $-1/1$  meaning perfect linear anticorrelation/correlation and values close to  $0$  meaning no linear correlation. Although Gradient Boosting Regressor has shown slightly better predicting performance and is less inclined to overfitting (scores for training of the model and evaluation are similar; see discussion below), we have decided to use Extra-Trees Regressor because it gives more consistent results between estimators (see below) and is computationally more efficient. This method works well on noisy data (Geurts et al. 2006).

Extra-Trees Regressor is an ensemble learning method that constructs multiple decision trees during training and outputs a mean prediction of the individual trees. This algorithm builds an ensemble of regression trees according to the classical top-down procedure. Two main differences with other tree-based ensemble methods are that it splits nodes using random thresholds for each feature rather than searching for the best possible thresholds and that it utilizes the whole learning sample (compared to a bootstrap replica, namely, resampling a data set with replacement) to grow the trees (Geurts et al. 2006; Geron 2019). We use Extra-Trees Regressor implemented in Scikit-Learn function `ExtraTreesRegressor` version `0.22.1`.

#### 3.1. Training the Model

The XMM-Newton SP count rates data set consists of data from 2001 January 2 to 2012 August 30. We took the data for training of the model and its validation from 2001 January 2 to 2010 December 31. The rest of the data from 2011 January 1 to 2012 August 30 are used only for the testing of the model. The



**Figure 5.** Correlation matrix between parameters. Here we used the Pearson correlation. The correlations are rounded to the second decimal for better visualization.

**Table 1**  
Performance of Different Models with Default Input Values

Regressor	Train Spearman	Test Spearman
Extra-Trees	1.000	0.441
Random Forest	0.947	0.402
Gradient Boosting	0.565	0.452
Multi-layer Perceptron	0.605	0.439
Stochastic Gradient Descent	0.447	0.408

ratio between amount of data for training/validation and testing is about 10:1.8. This is standard partitioning in ML (Geron 2019).

The parameters we want to correlate with the SP count rates all have different ranges; therefore, we decided to scale the data. We tried Scikit-Learn functions such as `StandardScaler`, `RobustScaler`, `MinMaxScaler`, and `QuantileTransformer`. The last scaler gave the best performance and is used in our model.

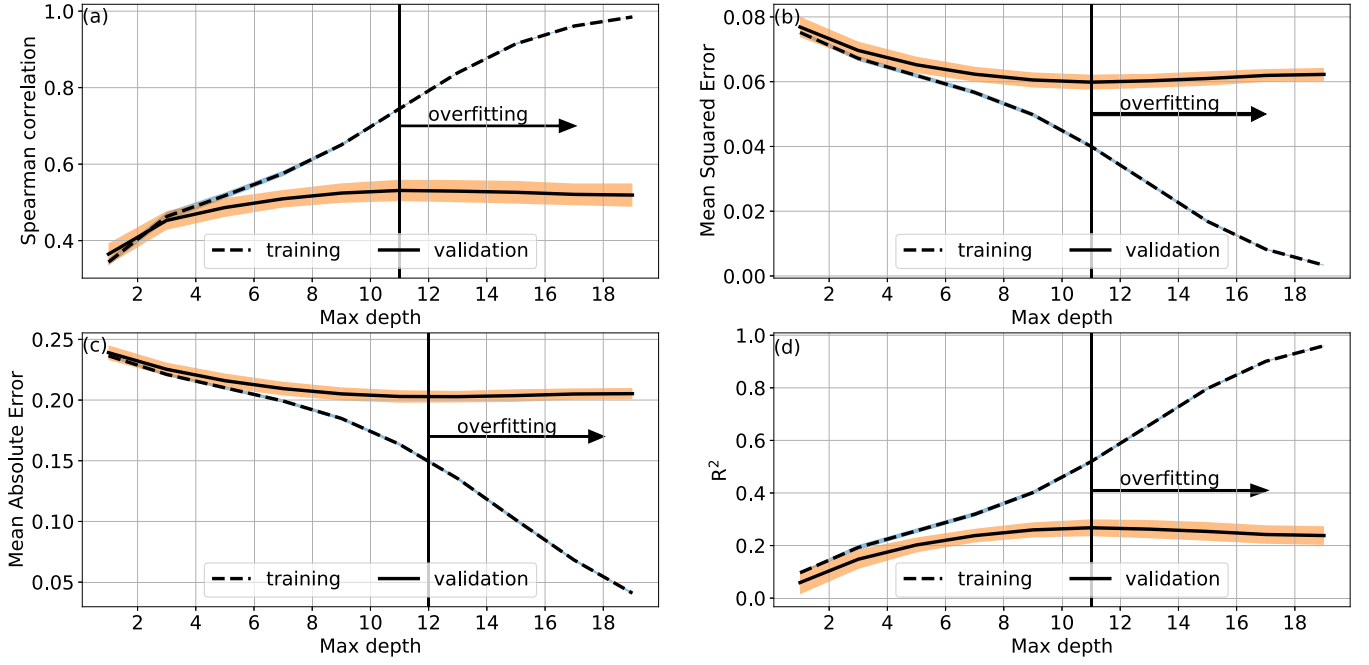
We trained the model using  $K$ -Folds cross-validation (function `model_selection.KFold`) with number of splits equal to 5. This method randomly divides the data set into  $K$  subsets of approximately the same size called folds, in our case  $K = 5$ . Then, we train and evaluate the Extra-Trees Regressor model five times by choosing a different fold for evaluation every time and training on the other four folds. This results in five arrays of evaluation scores. The advantage of training the model several times is that we can derive average

performance of the model for the train/validation data set, considering that in our case we observe a wide dynamic range in SP count rates. Another advantage is that one can estimate the precision of the model by deriving, e.g., its standard deviation.

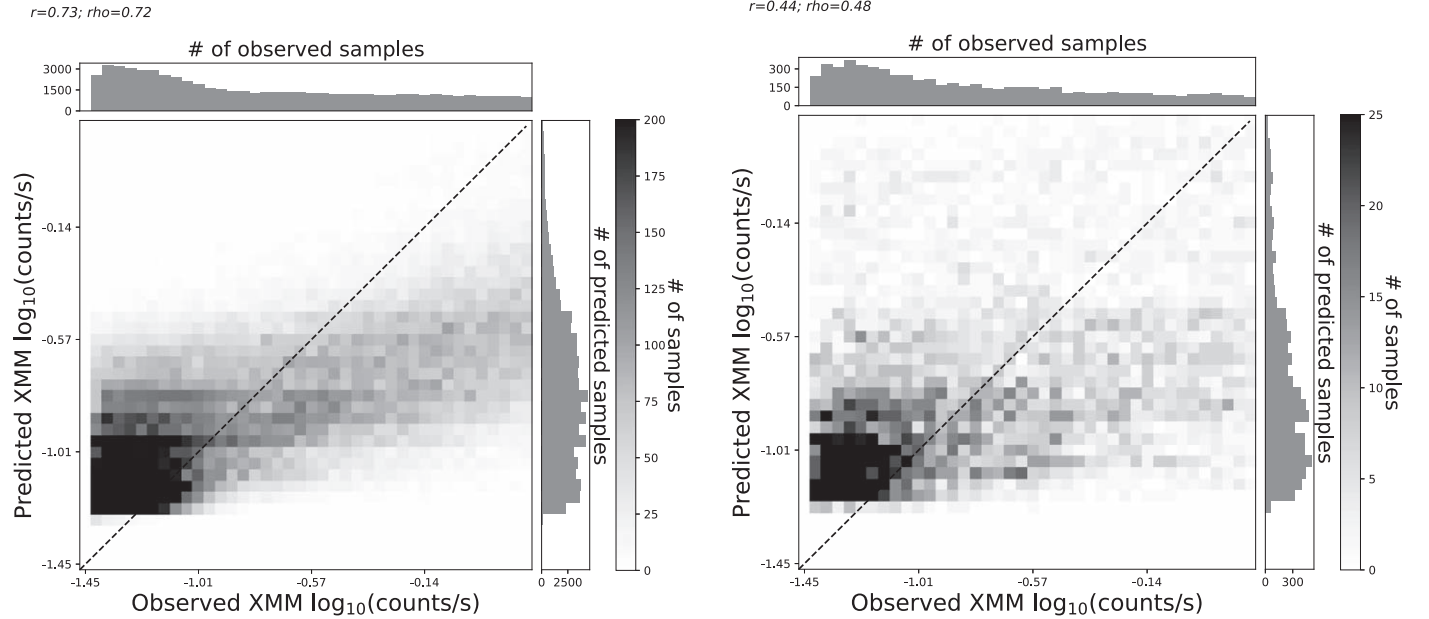
We use 1–200 trees with depths in the range from 1 to 20. Other parameters in the `ExtraTreesRegressor` were set as default. To evaluate the performance of the training and validation during cross-validation for different parameters, we use four different assessment metrics: Spearman correlation ( $\rho$ ), mean square error (MSE), mean absolute error (MAE), and coefficient of determination ( $R^2$ ). The values of MSE and MAE tend to zero in the case of perfect agreement between the model and observations.  $R^2$  indicates which fraction of data variability the model can explain; in the perfect case it is equal to 1.

To select best parameters of the estimator, we used optimization by cross-validated grid search over a parameter grid, `GridSearchCV`. This was done for four different metrics:  $\rho$ , MSE, MAE, and  $R^2$ . The performance of the model for the training/validation data sets is consistent between different metrics. The highest performance is observed for  $\approx 130$  trees and depth of 12 for MAE and 11 for MSE,  $\rho$ , and  $R^2$ . We plot the performance metrics of the model versus depth of the trees for the training and validation data sets for 130 trees in Figure 6. In the figure by a vertical line we indicate the optimal depth of the trees when a metric shows the minimum of validation error. For the depths of the trees with higher values the model starts to overfit the data (Prechelt 1998). Namely, the





**Figure 6.** Performance of the model for metrics  $\rho$ , MSE, MAE, and  $R^2$  (panels (a)–(d)) vs. depth of the trees for averaged training (solid line) and validation (dashed line) data sets. The number of estimators is equal to 130. The blue and orange colors indicate standard deviation for five cross-validation evaluation scores.



**Figure 7.** Observed count rates from the trained (left) and test (right) data set vs. those predicted by the model. The color represents number of samples.

discrepancy in performance between the training and validation data sets becomes larger (e.g., Ghogh & Crowley 2019). In the ideal case the gap between training and validation errors should be small (Goodfellow et al. 2017). For the model we select the depth of the trees equal to 11, the value at which the approximate minimum of validation error is observed. At this value the gap between training and validation errors is not too large yet.

The model is stable to outliers. We tried to limit the ranges of parameters; however, this did not improve the performance of the model significantly. We, therefore, do not limit the ranges of predictors.

The distribution of the observed count rates versus predicted by the model based on trained data set is shown in Figure 7 (left) and will be discussed in Section 3.3. The performance of the trained model evaluated by different estimators is listed in Table 2.

### 3.2. Predictor Importance

This method provides an opportunity to assess the relative importance of a feature with respect to the predictability of the target variable. This corresponds to the relative rank (tree depth) of a predictor used as a decision node in a tree. Features at the top of the tree affect the final prediction decision of a larger number of input samples. The relative importance of the

**Table 2**

Performance of the ML and Linear Models for Trained/Validation and Test Data Sets

Data Set	$r$	$\rho$	MSE	MAE	$R^2$
ML train	0.72	0.72	0.04	0.17	0.49
ML test	0.47	0.48	0.06	0.2	0.18
Linear Z train data set	0.35	0.32	0.45	0.57	0.12
Linear Z test data set	0.28	0.26	0.41	0.54	0.03
Linear $V_x$ train data set	0.19	0.20	0.42	0.55	0.01
Linear $V_x$ test data set	0.18	0.16	0.42	0.55	0.02

predictors is the expected fraction of the samples they affect. One can average the estimates of predictive ability over several randomized trees. This will reduce the variance of such an estimate and is called the mean decrease in impurity. In Scikit-Learn, the relative importance is combined with the mean decrease in impurity forming a normalized estimate of the predictive power of that feature (Pedregosa et al. 2011; Louppe 2014).

The relative importance is stored as an output in the fitted regression model. This is an array with shape corresponding to the number of features. The values of the array are positive and sum to 1.0. The higher the value, the more important is the contribution of the feature to the regression model. The relative importance of the features is plotted in Figure 8. The relative importance predicted by the Extra-Trees Regressor algorithm is consistent with Pearson correlations in Figure 5 and relations demonstrated in Figure 4: the location of the satellite, especially in the Z direction, the radial SW velocity, and the Foot Type are the most important parameters for the prediction of the contamination count rates.

We note that there are also other approaches to estimate feature importance such as Shapley values and permutation methods (Shapley 1953; Breiman 2001). Consideration of these methods, however, is beyond the scope of this work. The physics associated with important parameters is discussed in Section 4.

### 3.3. Testing the Model

We test the model on the available data from 2011 to 2012. The diagram of the model performance is shown in Figure 7 (right). The distribution of the observed and predicted counts indicates that the model (on both test and trained data sets; see Figure 7) underestimates high values and overestimates low values. This is also seen well in Figure 9, which presents the model performance on the test data set for a time interval in 2011. The performance of the trained model on the whole test data set, evaluated by different estimators, is listed in Table 2. The MSE error is close to zero. This is consistent with the ability of the model to predict mean values of count rates. The  $R^2$  indicates that the model can explain about 20% variability of the count rates. The Pearson and Spearman correlation coefficients are moderate and are statistically significant. To be significant at  $t = 0.05$  level, where  $t$  is the Student coefficient, the Pearson coefficient,  $r$ , needs to exceed the value defined by the following expression:

$$r = \frac{t}{n - 2 + t^2},$$

where  $n$  is the number of samples (Kendall & Stuart 1973). In our case taking  $t = 2.8$  for the two-sided distribution and

$n = 7341$ , the number of values in the test data set, we get that values of  $r > 0.03$  will give significant correlation. Our values of  $r$  are much higher, implying the significance of the model.

### 3.4. Comparison with Models Based on Best Individual Predictors

Performance of the ML model is moderate and is better than those of linear models derived with best individual predictors. In Figure 10 one can see the performance of the linear models using the training data set. The models not only strongly underestimate and overestimate high and low values, respectively, but also have lower performance estimated by Pearson and Spearman correlations, which are 0.35 and 0.32 for the model on ZGSE and 0.19 and 0.2 for the model on  $V_x$ SW\_GSE and by other metrics; see Table 2.

Less-than-optimal performance of the ML model can be explained by strongly nonlinear behavior of the energetic charged particles that trigger the contamination on short timescales. The performance of our ML model can be improved in the future by adding more data. The current data set covers just about one solar cycle and contains many data gaps (see Figure 9). However, the complete magnetic cycle of the Sun spans two solar cycles. This could introduce further variations, which we do not cover.

## 4. Discussion: Delineating Physics behind Contamination

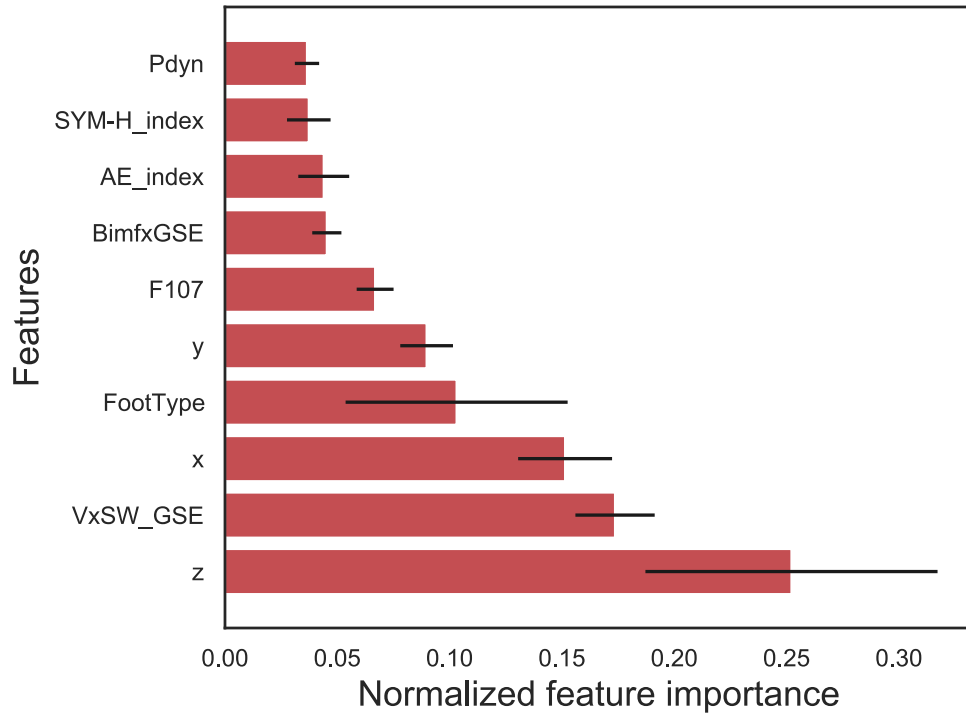
### 4.1. Spatial Dependencies of Contamination

The dependence of the SP contamination count rate on the spacecraft position in GSE coordinates is in agreement with previous results, e.g., by Kuntz & Snowden (2008) and Ghizzardi et al. (2017). Our results also agree with those from Walsh et al. (2014) that show the strongest contamination observed on the closed field lines (see Figure 2). The closed field lines are associated with the plasma sheet and the trapped particle population in the ring current and radiation belts. These are main reservoirs of energy in the magnetosphere. At higher latitudes, regions with open field lines or IMF (see Figure 2) become more important, leading to the decrease in the ZGSE direction. These regions are typically associated with particle energies well below the SP range. Indeed, SP count rates in the IMF are significantly lower than on closed field lines. Those on open magnetic field lines show the weakest SP count rates. The plasma on the IMF can be accelerated by shock related processes discussed in Sections 4.2.1 and 4.2.3.

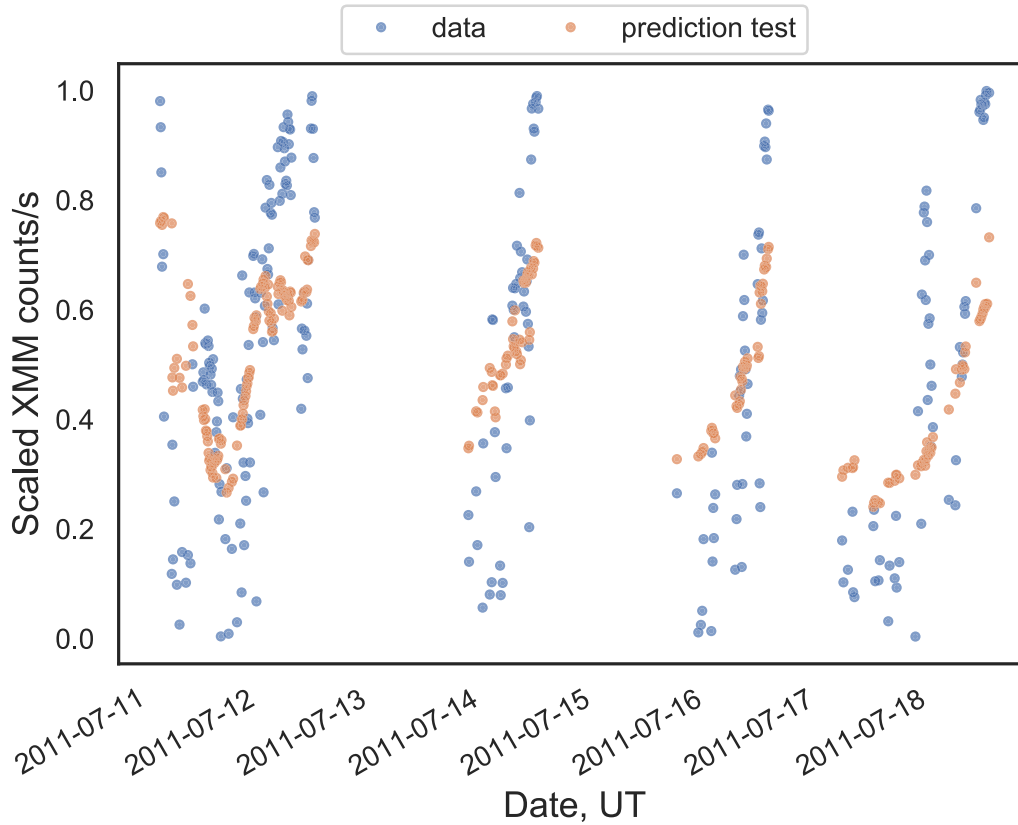
The duskward asymmetry of the contamination can partially be explained by loss of energetic particles toward the dawn side because of different loss mechanisms (see, e.g., Kronberg et al. 2014). Such asymmetry is observed for energetic protons ( $>274$  keV) and even stronger for energetic oxygen (see, e.g., Kronberg et al. 2015).

#### 4.1.1. Influence of IMF Direction

The general direction of the SW Parker spiral (Parker 1958) toward the Sun–Earth line,  $\phi$ , is  $\sim 45^\circ$ . In our data set, the average IMF components and confidence intervals are  $Bimfx_{GSE} = 0.018 \pm 0.025$  nT,  $Bimfy_{GSE} = -0.05 \pm 0.03$  nT, and  $Bimfz_{GSE} = 0.072 \pm 0.023$  nT. The average direction of the Parker spiral is  $\phi \sim 44^\circ$ . This geometry leads to formation of a quasi-parallel bow shock at the dawn side and a



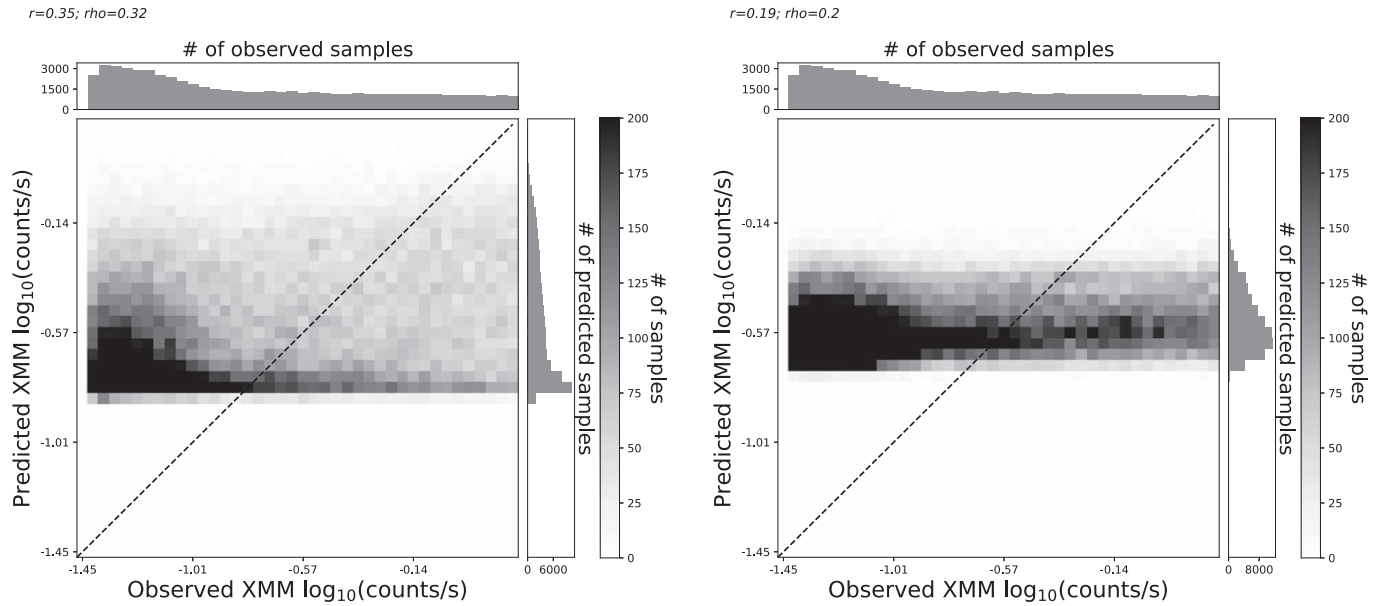
**Figure 8.** Importance of the different parameters in prediction of the contaminating count rates based on training data set. The black horizontal lines represent standard deviations.



**Figure 9.** Time profiles of an interval from the test data set of observed data and those predicted by the ML model.

quasi-perpendicular bow shock at the dusk side. The quasi-parallel bow shocks are strong accelerators of plasma; see Section 4.2.3. Therefore, it would be expected to observe more

contamination at the dawn side. However, this is not observed, and consequently the direction of Parker spiral cannot explain the duskward asymmetry.



**Figure 10.** Observed count rates from the trained data set vs. those predicted by the linear models, depending on ZGSE (left) and  $V_{xSW\_GSE}$  (right). The color represents number of samples.

At higher latitudes, the dayside and duskward flank preference of the contamination (see Figure 4(c) for the dayside asymmetry and Figure 3 for the duskward asymmetry) can be explained by the location of acceleration sources for particles at the day side and the location of reconnection (e.g., Luo et al. 2017). In this study, the asymmetries in the spatial distributions of energetic ions were related to the location of the reconnection. In the case of an average downward and northward IMF direction, the reconnection location is expected at dusk side high latitudes in the southern hemisphere (Luo et al. 2017). Particles on the reconnected field lines are further accelerated, e.g., in a diamagnetic cavity (a region with low magnetic field formed during reconnection close to cusp (see sketch in Figures 2 and 6 in Nykyri et al. 2011)). This region traps and accelerates plasma particle population that then can penetrate inside the magnetosphere and populate it (e.g., Nykyri et al. 2012; Sorathia et al. 2019). In the data, the direction of the IMF is slightly northward and downward. However, the errors introduced by the processing of the OMNI data are in the range of 0.2 nT (King & Papitashvili 2005). Therefore, we cannot statistically confirm this explanation. More work is needed in this direction.

#### 4.2. Dependence on SW Velocity

Significant growth of the logarithm of the count rates with increasing radial SW velocity,  $V_x$ , means that  $V_x$  can be considered as a most important space weather parameter related to the XMM-Newton contamination. The increase of the SW speed leads to the compression of the magnetosphere. However, the count rates show rather weak dependence on the SW dynamic pressure; see Figure 4(h). Therefore, additional processes associated with faster SW lead to enhanced contamination.

##### 4.2.1. Influence of SW High-speed Streams

High SW speed is often associated with the SW high-speed streams (HSSs). HSSs mostly occur during declining phase of

the solar cycle owing to an increase in equatorial coronal holes, which are the source of HSSs. In Figure 4(j) one can see that the strongest contamination occurs during medium values of the parameter F10.7, which corresponds to the declining phase of the solar cycle. Regions in which HSSs overtake slow SW are often associated with CIRs. At larger heliospheric distances (beyond 1 astronomical unit (au)), CIRs often form shocks, which accelerate ions (e.g., Richardson 2018). These accelerated ions can then travel back toward the Sun and have been observed within about 0.3 au (e.g., Allen et al. 2020). Both the abundance and the composition of suprathermal ions associated with CIRs have also been shown to have a solar cycle dependence (e.g., Allen et al. 2019). These accelerated ions can enter into the magnetosphere via reconnection.

##### 4.2.2. Influence of SW–Magnetosphere Energy Coupling

The SW speed is proportional to the SW electric field that controls the magnetic reconnection rate at the day side (Dorelli 2019). Increased SW speed will lead to increased reconnection rate. Most of the SW–magnetosphere energy coupling functions are proportional to the SW speed (e.g., Gonzalez et al. 1994; Milan et al. 2012; Wang et al. 2014). The increase in the SW speed leads to more effective further transport of the reconnected magnetic field lines toward the tail and then, again via reconnection on the night side, back to Earth to complete the cycle. A brief disturbance ( $\sim 3$  hr) that causes energy release from the tail into the high-latitude ionosphere is called a substorm. This will lead to deviation of the magnetic field on the ground in the high-latitude regions and will be reflected in the AE index. This agrees well with the AE index being in the set of parameters leading to better prediction of the ML model; see Figure 8(g). Significant growth of the count rates with AE index at least up to 100 nT is observed. Substorm activity leads to strong acceleration of ions by processes associated with magnetic reconnection such as magnetic field dipolarization (see, e.g., Grigorenko et al. 2017). Stronger substorm activity does not lead to more effective



acceleration of ions (see the same result in Luo et al. 2014). This is probably related to more effective loss mechanisms of particles producing SPs during high magnetospheric activity. For example, acceleration to higher energy can lead to a decrease in the SP population. This is a topic for future investigations. We also would like to note that AE index is measured in the northern hemisphere, although XMM-Newton observations are in the southern hemisphere. This fact may reduce correlation between observations of count rates and northern geomagnetic activity. Weygand & Zesta (2008) have shown that observations of southern auroral region ground magnetometers are not always consistent with AE index.

Increased SW speed on longer timescales (hours), e.g., during CMEs, may lead to geomagnetic storms. The SP count rates increase with decrease of the SYM-H index from 0 up to approximately  $-50$  nT; see Figure 4(l). At stronger magnetic storms nonlinear behavior is observed. The same as strong substorms, strong magnetic storms can be associated with higher losses of SPs. Such nonlinear behavior of SYM-H index and AE index indicates that alone they are not necessarily good parameters for prediction of the XMM-Newton contamination during geomagnetically active times.

#### 4.2.3. Role of Quasi-parallel Bow Shock

In Figure 4(i) one can see an increase of the contamination during large absolute values of the IMF  $B_x$  component. Large absolute values of the IMF  $B_x$  will increase probability of the formation of the quasi-parallel bow shock (normal to the shock is parallel to the IMF direction), at least at the dayside magnetosphere. The quasi-parallel bow shocks are strong accelerators of plasma (e.g., Blandford & Ostriker 1978; Kronberg et al. 2009; Sundberg et al. 2016). The shocks with higher Mach numbers, associated with higher SW speeds, lead to more effective ion acceleration (Treumann 2009).

#### 4.3. Oxygen Ions

We compare the dependence of the SP count rates on the AE index and the SW dynamic pressure with the dependencies of proton and oxygen ions at 10 keV and  $>274$  keV in the terrestrial plasma sheet observed by Kronberg et al. (2012). One can note that the trends in Figures 4(h) and (k) are similar to the dependence of energetic hydrogen and oxygen ions ( $>274$  keV) on the AE index and the SW dynamic pressure in that study. This is consistent with the idea that energetic protons at several hundreds of keV may produce contamination. Additionally, this indicates that oxygen ions may also contaminate the XMM-Newton telescope. Kronberg et al. (2012) show that the intensity of oxygen ions can be comparable to those of protons during disturbed magnetospheric activity.

### 5. Conclusions and Open Questions

In this paper we delineate which geometric, solar, SW, and geomagnetic parameters are related to strong contamination in the XMM-Newton telescope, derive prediction models, and discuss the possible physical interpretation suggested by this approach.

1. We reveal strong association of the contamination with (a) location of the satellite and, therefore, the region in space (the strongest and clear exponential dependence is

derived for the southward direction,  $Z$ ); (b) the radial SW speed (exponential dependence is derived); and (c) magnetic field line Foot Type (the strongest contamination is observed on closed field lines).

2. We derived a model to predict contamination that utilizes an ensemble of predictors (Extra-Trees Regressor). It shows better performance than models based on individual parameters such as  $Z$  or  $V_x$ . It also helps to quantify importance for nonlinear relations.
3. The analysis of relative importance of the parameters indicates that (a) processes of acceleration related to formation of the quasi-parallel shock may play an important role in formation of the contaminating population, indications for these being (i) relatively strong contamination at large absolute values of IMF  $B_x$ , (ii) strong dependence on the SW velocity, and (iii) stronger contamination at the day side; (b) acceleration processes associated with reconnection at the day side may also play an important role; and (c) SYM-H index and AE index alone are not necessarily good parameters for prediction of the XMM-Newton contamination during geomagnetically active times.
4. Similarity of the dependencies of the SP count rates and the energetic oxygen ( $>274$  keV) in the plasma sheet on the AE index and the SW dynamic pressure gives a hint that oxygen may also contaminate the XMM-Newton telescope.

Road map for future missions: (a) it is advisable to avoid observations during times associated with high SW speed in the near-Earth magnetospheric region, and (b) the same is recommended for closed magnetic field lines, especially at the dusk flank in the southern hemisphere (asymmetries in the northern hemisphere are not studied here).

In our next studies we will focus on the following questions: (1) Which processes associated with the strong SW speed are effective accelerators of energetic particles? In particular, acceleration sources associated with reconnection at the day side (such as diamagnetic cavities at cusps) and quasi-parallel bow shocks require enhanced attention. (2) Which energy ranges of particles are most efficient at producing this contamination? (3) Are there losses of SP contaminating particles during high magnetospheric activities? (3) What role do energetic oxygen ions play in the contamination observed by XMM? We will address all these questions in future work. For this we plan to compare XMM-Newton observations with energetic particle observations by the Cluster mission.

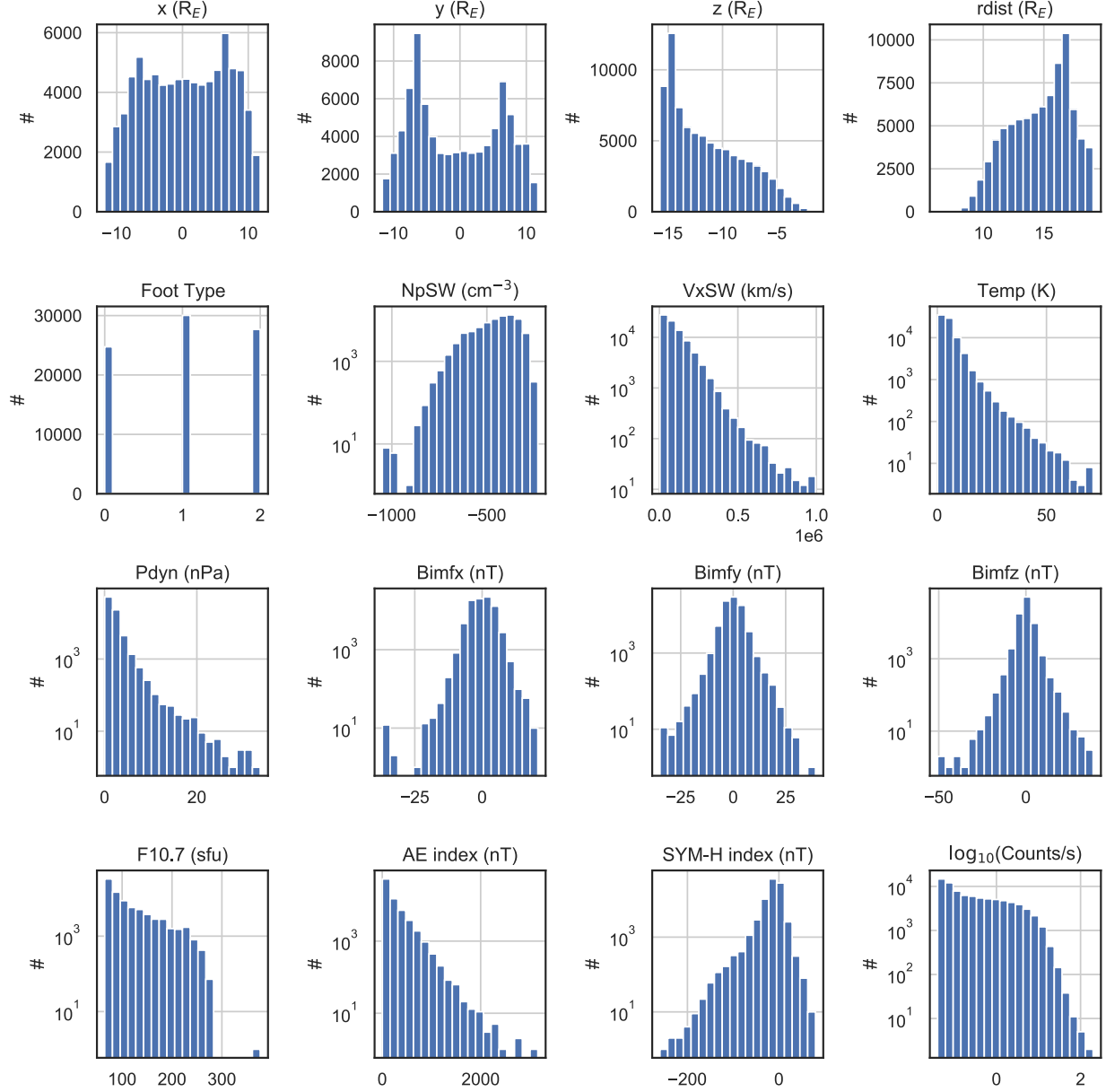
We acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb service and OMNI data. We acknowledge XMM-Newton data archive <https://www.cosmos.esa.int/web/xmm-newton/xsa>. The AE and SYM-H indices used were provided by the WDC for Geomagnetism, Kyoto (<http://wdc.kugi.kyoto-u.ac.jp/wdc/Sec3.html>). F10.7 index can be found at <https://spaceweather.gc.ca/solarflux/sx-5-en.php>. This work was conceived within the team led by Fabio Gastaldello on "Soft Protons in the Magnetosphere focused by X-ray Telescopes" at the International Space Science Institute in Bern, Switzerland. N.S. is supported by NASA Earth and Space Science grant 80NSSC17K0433. E.K. is supported by German Research Foundation (DFG) under No. KR 4375/2-1 within SPP "Dynamic Earth." The work of E.K. and A.S. is

supported by the Volkswagen Foundation grant Az 90 312. We are thankful to Irina Zhelavskaya for advice related to ML.

*Software:* sklearn (Pedregosa et al. 2011), scipy (Virtanen et al. 2020), numpy (van der Walt et al. 2011), pandas (McKinney 2010), Matplotlib (Hunter 2007), mysql.connector.














## Appendix

Figure A1 shows the distribution of the number of samples for the predictors and the count rates (on the vertical axis) with a given value range (on the horizontal axis).



**Figure A1.** Histograms of the number of samples of predictors and SP count rates.

## ORCID iDs

Elena A. Kronberg  <https://orcid.org/0000-0001-7741-682X>  
 Fabio Gastaldello  <https://orcid.org/0000-0002-9112-0184>  
 Stein Haaland  <https://orcid.org/0000-0002-1241-7570>  
 Artem Smirnov  <https://orcid.org/0000-0003-3689-4336>  
 Max Berrendorf  <https://orcid.org/0000-0001-9724-4009>  
 Simona Ghizzardi  <https://orcid.org/0000-0003-0879-7328>  
 K. D. Kuntz  <https://orcid.org/0000-0001-6654-5378>  
 Nithin Sivadas  <https://orcid.org/0000-0003-4278-0482>  
 Robert C. Allen  <https://orcid.org/0000-0003-2079-5683>  
 Andrea Tiengo  <https://orcid.org/0000-0002-6038-1090>  
 Raluca Ilie  <https://orcid.org/0000-0002-7305-2579>  
 Yu Huang  <https://orcid.org/0000-0001-5023-0427>  
 Lynn Kistler  <https://orcid.org/0000-0002-8240-5559>

## References

- Allen, R. C., Ho, G. C., & Mason, G. M. 2019, *ApJL*, **883**, L10  
 Allen, R. C., Lario, D., Odstrcil, D., et al. 2020, *ApJS*, **246**, 36  
 Blandford, R. D., & Ostriker, J. P. 1978, *ApJL*, **221**, L29  
 Breiman, L. 2001, *Machine Learning*, **45**, 5  
 Camporeale, E. 2019, *SpWea*, **17**, 1166  
 Carter, J. A., & Read, A. M. 2007, *A&A*, **464**, 1155  
 Chu, X., Bortnik, J., Li, W., et al. 2017, *JGRA*, **122**, 9183  
 De Luca, A., & Molendi, S. 2004, *A&A*, **419**, 837  
 De Luca, A., Salvaterra, R., Tiengo, A., et al. 2017, in *The X-ray Universe 2017 Symp., EXTraS: Exploring the X-ray Transient and variable Sky*, ed. J.-U. Ness & S. Migliari (Paris: ESA), **65**  
 Dorelli, J. C. 2019, *JGRA*, **124**, 2668  
 Escoubet, C. P., Schmidt, R., & Goldstein, M. L. 1997, *SSRv*, **79**, 11  
 Fioretti, V., Bulgarelli, A., Malaguti, G., Spiga, D., & Tiengo, A. 2016, *Proc. SPIE*, **9905**, 99056W  
 Fioretti, V., Bulgarelli, A., Molendi, S., et al. 2018, *ApJ*, **867**, 9  
 Gastaldello, F., Ghizzardi, S., Marelli, M., et al. 2017, *ExA*, **44**, 321  
 Geron, A. 2019, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (Sebastopol, CA: O'Reilly & Associates)  
 Geurts, P., Ernst, D., & Wehenkel, L. 2006, *Machine Learning*, **63**, 3  
 Ghizzardi, S., Marelli, M., Salvetti, D., et al. 2017, *ExA*, **44**, 273  
 Ghogh, B., & Crowley, M. 2019, arXiv:1905.12787  
 Gonzalez, W. D., Joselyn, J. A., Kamide, Y., et al. 1994, *JGR*, **99**, 5771  
 Gonzalez, W. D., Tsurutani, B. T., & Clúa de Gonzalez, A. L. 1999, *SSRv*, **88**, 529  
 Goodfellow, I., Bengio, Y., & Courville, A. 2017, *Machine Learning Basics* (Cambridge, MA: MIT Press), 98  
 Grant, C. E., Ford, P. G., Bautz, M. W., & O'Dell, S. L. 2012, *Proc. SPIE*, **8443**, 844311  
 Grigorenko, E. E., Kronberg, E. A., & Daly, P. W. 2017, *CosRe*, **55**, 57  
 Hunter, J. D. 2007, *CSE*, **9**, 90  
 Jansen, F., Lumb, D., Altieri, B., et al. 2001, *A&A*, **365**, L1  
 Kendall, M. G., & Stuart, A. 1973, *The Advanced Theory of Statistics: Inference and Relationship* (Griffin)  
 King, J. H., & Papitashvili, N. E. 2005, *JGRA*, **110**, A02104  
 Kronberg, E. A., Ashour-Abdalla, M., Dandouras, I., et al. 2014, *SSRv*, **184**, 173  
 Kronberg, E. A., Grigorenko, E. E., Haaland, S. E., et al. 2015, *JGRA*, **120**, 3415  
 Kronberg, E. A., Haaland, S. E., Daly, P. W., et al. 2012, *JGR*, **117**, 12208  
 Kronberg, E. A., Kis, A., Klecker, B., Daly, P. W., & Lucek, E. A. 2009, *JGR*, **114**, 3211  
 Kuntz, K. D., & Snowden, S. L. 2008, *A&A*, **478**, 575  
 Laurenza, M., Alberti, T., Marcucci, M. F., et al. 2019, *ApJ*, **873**, 112  
 Li, Y., Luhmann, J. G., & Lynch, B. J. 2018, *SoPh*, **293**, 135  
 Lotti, S., Mineo, T., Jacquey, C., et al. 2018, *ExA*, **45**, 411  
 Louppe, G. 2014, PhD thesis, Univ. Liege  
 Luo, H., Kronberg, E. A., Grigorenko, E. E., et al. 2014, *GeoRL*, **41**, 3724  
 Luo, H., Kronberg, E. A., Nykyri, K., et al. 2017, *JGRA*, **122**, 5168  
 Marelli, M., Salvetti, D., Gastaldello, F., et al. 2017, *ExA*, **44**, 297  
 McKinney, W. 2010, in *Proc. 9th Python in Science Conf., Data Structures for Statistical Computing in Python*, ed. S. van der Walt & J. Millman (Austin, TX: SciPy), 56  
 McPherron, R. L., & Weygand, J. 2006, in *Recurrent Magnetic Storms: Corotating Solar Wind*, ed. B. Tsurutani et al., Vol. 167 (Washington, DC: AGU), 125  
 Milan, S. E., Gosling, J. S., & Hubert, B. 2012, *JGRA*, **117**, A03226  
 Nandra, K., Barret, D., Barcons, X., et al. 2013, arXiv:1306.2307  
 Nose, M., Sugiura, M., Kamei, T., & Iyemori, T. 2017, *AE Index, WDC for Geomagnetism, Kyoto*, doi:10.17593/15031-54800  
 Nykyri, K., Otto, A., Adamson, E., Dougal, E., & Mumme, J. 2011, *JGRA*, **116**, A03228  
 Nykyri, K., Otto, A., Adamson, E., Kronberg, E., & Daly, P. 2012, *JASTP*, **87**, 70  
 O'Dell, S. L., Bautz, M. W., Blackwell, W. C., et al. 2000, *Proc. SPIE*, **4140**, 99  
 O'Dell, S. L., Blackwell, W. C. J., Cameron, R. A., et al. 2003, *Proc. SPIE*, **4851**, 77  
 Parker, E. N. 1958, *ApJ*, **128**, 664  
 Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *Journal of Machine Learning Research*, **12**, 2825  
 Prechelt, L. 1998, *Early Stopping—But When?* (Berlin: Springer), 55  
 Prigozhin, G. Y., Kissel, S. E., Bautz, M. W., et al. 2000a, *Proc. SPIE*, **4012**, 720  
 Prigozhin, G. Y., Kissel, S. E., Bautz, M. W., et al. 2000b, *Proc. SPIE*, **4140**, 123  
 Richardson, I. G. 2018, *LRSP*, **15**, 1  
 Roberts, D. A., Karimabadi, H., Sipes, T., Ko, Y.-K., & Lepri, S. 2020, *ApJ*, **889**, 153  
 Salvetti, D., Marelli, M., Gastaldello, F., et al. 2017, *ExA*, **44**, 309  
 Shapley, L. S. 1953, *A Value for n-person Games*, 307  
 Smirnov, A. G., Berrendorf, M., Shprits, Y. Y., et al. 2020, *SpWea*, **18**, e2020SW002532  
 Sorathia, K. A., Merkin, V. G., Ukhorskiy, A. Y., et al. 2019, *JGRA*, **124**, 5461  
 Strüder, L., Briel, U., Dennerl, K., et al. 2001, *A&A*, **365**, L18  
 Sundberg, T., Haynes, C. T., Burgess, D., & Mazelle, C. X. 2016, *ApJ*, **820**, 21  
 Tapping, K. F. 2013, *SpWea*, **11**, 394  
 Treumann, R. A. 2009, *A&ARv*, **17**, 409  
 Tsyganenko, N. A. 1995, *JGR*, **100**, 5599  
 Turner, M. J. L., Abbey, A., Arnaud, M., et al. 2001, *A&A*, **365**, L27  
 van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, *CSE*, **13**, 22  
 Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nature Methods*, **17**, 261  
 Walsh, B. M., Kuntz, K. D., Collier, M. R., et al. 2014, *SpWea*, **12**, 387  
 Wang, C., Han, J. P., Li, H., Peng, Z., & Richardson, J. D. 2014, *JGRA*, **119**, 6199  
 Weisskopf, M. C., Brinkman, B., Canizares, C., et al. 2002, *PASP*, **114**, 1  
 Weygand, J. M., & Zesta, E. 2008, *JGRA*, **113**, A08202  
 Zhelavskaya, I. S., Shprits, Y. Y., & Spasojević, M. 2017, *JGRA*, **122**, 227