Statistical models of the variability of plasma in the topside ionosphere

Wood, Alan G.; Donegan-Lawley, Elizabeth E.; Clausen, Lasse B. N.; Spogli, Luca; Urbář, Jaroslav; Jin, Yaqi; Shahtahmassebi, Golnass; Alfonsi, Lucilla; Rawlings, James T.; Cicone, Antonio; Kotova, Daria; Cesaroni, Claudio; Høeg, Per; Dorrian, Gareth D.; Nugent, Luke D.; Elvidge, Sean; Themens, David R.; Aragón, María José Brazal; Wojtkiewicz, Pawel; Miloch, Wojciech J.

DOI:
10.1051/swsc/2024002

J. Space Weather Space Clim. 2024, **14**, 7

https://doi.org/10.1051/swsc/2024002

JSWSC

Available online at:
www.swsc-journal.org

RESEARCH ARTICLE OPEN ACCESS

Statistical models of the variability of plasma in the topside ionosphere: 1. Development and optimisation

Alan G. Wood[1,*], Elizabeth E. Donegan-Lawley[1], Lasse B. N. Clausen[2], Luca Spogli[3,4], Jaroslav Urbář[3,5], Yaqi Jin[2], Golnaz Shahtahmassebi[6], Lucilla Alfonsi[3], James T. Rawlings[6], Antonio Cicone[3,7], Daria Kotova[2], Claudio Cesaroni[3], Per Høeg[2], Gareth D. Dorrian[1], Luke D. Nugent[1], Sean Elvidge[1], David R. Themens[1], María José Brazal Aragón[8], Pawel Wojtkiewicz[8], and Wojciech J. Miloch[2]

[1] Space Environment and Radio Engineering (SERENE) group, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
[2] Department of Physics, University of Oslo, PO Box 1048, Blindern, 0316 Oslo, Norway
[3] Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata, 605, 00143 Rome, Italy
[4] SpacEarth Technology, Via di Vigna Murata, 605, 00143 Rome, Italy
[5] Institute of Atmospheric Physics of the Czech Academy of Sciences, Boční II 1401, 141 00 Prague, Czech Republic
[6] Department of Physics & Mathematics, Nottingham Trent University, Clifton Lane, Clifton, Nottingham NG11 8NS, UK
[7] Università Degli Studi Dell'Aquila, via Vetoio, 1, 67100 L'Aquila, Italy
[8] GMV Innovating Solutions, Hrubieszowska 2, 01-209 Warsaw, Poland

Received 3 May 2023 / Accepted 10 January 2024

Abstract – This work presents statistical models of the variability of plasma in the topside ionosphere based on observations made by the European Space Agency's (ESA) Swarm satellites. The models were developed in the "Swarm Variability of Ionospheric Plasma" (Swarm-VIP) project within the European Space Agency's Swarm+4D-Ionosphere framework. The configuration of the Swarm satellites, their near-polar orbits and the data products developed, enable studies of the spatial variability of the ionosphere at multiple scale sizes. The statistical modelling technique of Generalised Linear Modelling (GLM) was used to create models of both the electron density and measures of the variability of the plasma structures at horizontal spatial scales between 20 km and 100 km. Despite being developed using the Swarm data, the models provide predictions that are independent of these data. Separate models were created for low, middle, auroral and polar latitudes. The models make predictions based on heliogeophysical variables, which act as proxies for the solar and geomagnetic processes. The first and most significant term in the majority of the models was a proxy for solar activity. The most common second term varied with the latitudinal region. This was the Solar Zenith Angle (SZA) in the polar region, a measure of latitude in the auroral region, solar time in the mid-latitude region and a measure of latitude in the equatorial region. Other, less significant terms in the models covered a range of proxies for the solar wind, geomagnetic activity and location. In this paper, the formulation, optimisation and evaluation of these models are discussed. The models show very little bias, with a mean error of zero to two decimal places in 14 out of 20 cases. The models capture some, but not all, of the trends present in the data, with Pearson correlation coefficients of up to 0.75 between the observations and the model predictions. The models also capture some, but not all, of the variability of the ionospheric plasma, as indicated by the precision, which ranged between 0.20 and 0.83. The addition of the thermospheric density as an explanatory variable in the models improved the precision in the polar and auroral regions. It is suggested that, if the thermosphere could be observed at a higher spatial resolution, then even more of the variability of the plasma structures could be captured by statistical models. The formulation and optimisation of the models are presented in this paper. The capability of the model in reproducing the expected climatological features of the topside ionosphere, in supporting GNSS-based ionospheric observations and the performance of the model against the Thermosphere-Ionosphere-Electrodynamics General Circulation Model (TIE-GCM), are provided in a companion paper (Spogli L et al. 2024. J Space Weather Space Clim https://doi.org/10.1051/swsc/2024003).

Keywords: Topside ionosphere / Space weather / Ionosphere atmosphere interactions / Statistical modelling

*Corresponding author: a.wood.1@bham.ac.uk

1 Introduction

The F-region of the ionosphere is a highly complex plasma containing density structures with a wide range of spatial scales. Large-scale structures with horizontal extents of tens to hundreds of km exhibit variation with time of day, season, solar cycle, geomagnetic activity, solar wind conditions and location. Plasma is primarily created by ionisation of the upper atmosphere by solar extreme ultraviolet (EUV) radiation and it decays by recombination with neutral species in the atmosphere. The intensity of the incident solar radiation is a function of the solar zenith angle (SZA), therefore a diurnal and seasonal variation in the production rate of ionisation is expected (Pedersen, 1927). The solar EUV flux varies during the solar cycle (Hinteregger, 1977) and a variation in the production rate of ionospheric plasma is expected on these timescales. The bulk properties of the ionosphere are also influenced by the neutral atmosphere. Rishbeth & Setty (1961) and Wright (1963) reported that the ionospheric density was greater during winter than summer at mid-latitudes. This is known as the seasonal anomaly, also referred to as the winter anomaly. These authors attributed this effect to higher summer temperatures which caused upwelling of the thermosphere in the summer hemisphere. This led to lower O/N_2 and O/O_2 atomic/molecular concentration ratios, which increased the recombination rate and consequently decreased the plasma density. The ionosphere exhibits several other anomalies, which were summarised by Hargreaves (1992). These include the annual anomaly and semi-annual anomaly. The annual anomaly is that the global average plasma density is greater in December than in June by 20%. This can be partially explained by the annual variation in the Sun-Earth distance. The semi-annual anomaly is that the global average plasma density is greater at the equinox than at the solstice. This is attributed to the temperature gradient between the summer and winter poles at the solstice, driving winds that transport molecular-rich air from the summer to the winter pole, increasing the recombination rate of the plasma.

Plasma structures are commonly observed in the ionosphere. At equatorial and low latitudes, the equatorial ionospheric anomaly (EIA) arises due to the combined effects of the daytime equatorial electrojet and the terrestrial magnetic field. The EIA was first reported by Appleton (1946) and has been extensively characterised since, as reviewed by Balan et al. (2018). The decay of plasma by chemical recombination is faster at lower altitudes, due to the neutral atmosphere density profile. Therefore, after sunset, a steep vertical density gradient forms and plasma structures grow due to instability processes and the pre-reversal enhancement in equatorial vertical drift, driven by the equatorial electrojet and F-region dynamo winds. Plasma density irregularities are commonly observed in the low-latitude ionosphere after sunset (Kil & Heelis, 1998), which can be identified as plasma density depletions known as equatorial plasma bubbles (EPBs) (McClure et al., 1977). They affect radio signals, causing effects such as the range and frequency spread signatures on high-frequency (HF) echoes known as equatorial spread F (Woodman & La Hoz, 1976) and scintillation on VHF-UHF and L-band signals (Basu & Basu, 1981).

At high latitudes, polar cap patches are commonly observed. These were defined by Crowley (1996) to have a horizontal extent of at least 100 km and a plasma density of at least twice that of the surrounding background ionosphere. A polar cap patch was first reported by Hill (1963) and was observed to drift with the background plasma flow (Buchau et al., 1983). It was proposed that such patches were produced on the dayside at auroral or subauroral latitudes and then drawn into the polar cap by the high-latitude convection pattern (Weber et al., 1984). An individual patch was tracked for more than 3000 km (Weber et al., 1986). Patches have been observed to drift out of the polar cap (Pedersen et al., 2000) and to be reconfigured to form a boundary blob (Pryse et al., 2006; Jin et al., 2016). Polar cap plasma exhibits seasonal variation (Foster, 1984), but plasma structures can persist in summer even if they do not meet the formal definition of a polar cap patch (Wood & Pryse, 2010). Polar cap patches can derive from transient bursts of reconnection in the magnetosphere (Lockwood & Carlson, 1992), variations in the Interplanetary Magnetic Field (IMF) altering the source region of plasma drawn into the polar cap (Sojka et al., 1993), variations in the IMF determining whether this plasma can enter the polar cap (Valladares et al., 1998) or the fragmentation of the tongue of ionization (Rodger et al., 1994; Valladares et al., 1994; De Franceschi et al., 2008). Birkeland (1913) suggested that a stream of charged particles from the Sun could be guided by the geomagnetic field to impact the polar atmosphere and cause the aurora. The process of particle precipitation also results in the ionisation of the upper atmosphere (Rees, 1989; Brekke, 1997), which can result in the formation of plasma structures (Walker et al., 1999) known as "hot" patches (Zhang et al., 2017).

At mid-latitudes, plasma structures are observed, which have propagated latitudinally to this region from lower or higher latitudes (e.g. Fallows et al., 2020), or which result from vertical coupling from lower altitudes (Rishbeth & Mendillo, 2001). Travelling Ionospheric Disturbances (TIDs) are commonly observed at these latitudes. These are horizontally propagating waves which can result from auroral precipitation, heating from ionospheric current systems and atmospheric gravity waves propagating from the lower atmosphere, as reviewed by Hunsucker (1982). TIDs are observed or inferred at a wide range of scale sizes, with wavelengths ranging from of the order of 1000 km (Francis, 1975) to less than 30 km (Boyde et al. 2022). Fallows et al. (2020) simultaneously observed large and medium-scale TIDs in the mid-latitude ionosphere at different altitudes propagating horizontally and approximately perpendicular to each other. Cherniak & Zakharenkova (2016) and Cherniak et al. (2019) observed ionospheric plasma bubbles at mid-latitudes which had propagated from the equatorial region. Additionally, the atmospheric events induced by the eruption at Hunga Tonga-Hunga Ha'apai have consolidated the evidence about how natural hazards are major sources of TIDs affecting the mid-latitude ionosphere through Lithosphere-Atmosphere-Ionosphere coupling (e.g. Rajesh et al., 2022; Sun et al., 2022; Themens et al., 2022; Wright et al., 2022).

Plasma structures can cause challenges for trans-ionospheric radio signals. Variations in the plasma density result in changes to the refractive index of the ionosphere (Hargreaves, 1992). Trans-ionospheric radio waves undergo refraction and/or diffraction (Wernik et al., 2003). The interference of the scattered waves can result in rapid variations in the phase and intensity of the received signal, a phenomenon known as scintillation. This was first reported by Hey et al. (1946) who conducted radio astronomical observations of Cygnus-A at 64 MHz. Ionospheric

scintillation has become of increasing concern in recent years due to the increasing importance of practical navigation and communication systems, such as Global Navigation Satellite Systems (GNSS). A direct connection between gradients in the Total Electron Content at the edge of a plasma stream and scintillation has been observed (Mitchell et al., 2005) and plasma structuring caused by auroral precipitation has been linked to the loss of signal lock by a GNSS receiver (Smith et al., 2008; Elmas et al., 2011; Jin & Oksavik, 2018). Statistical studies have shown the climatology of ionospheric scintillation at GNSS frequencies (Prikryl et al., 2015), that auroral emissions correlate with GNSS signal scintillation (Kinrade et al., 2013), an agreement between scintillation and the expected position of the cusp and auroral oval boundaries, and between scintillation and large scale plasma structures including polar cap patches and EPBs (Spogli et al., 2009; Jin et al., 2014; De Franceschi et al., 2019; Li et al., 2021). Plasma structures can occur without scintillation (e.g. Jenner et al., 2020) and it has been suggested that both a minimum gradient in electron density and a minimum value of electron density are required for scintillation to occur (Aarons, 1982). The nature of scintillation and its connection with refractive and diffractive mechanisms causing the observed amplitude and phase fluctuations have been recently debated (see, e.g. McCaffrey & Jayachandran 2019; Ghobadi et al., 2020; Spogli et al., 2021).

Plasma structuring in the ionosphere can be successfully studied in situ with satellite missions, such as Swarm. Swarm is the European Space Agency's (ESA) first constellation mission for Earth Observation (Friis-Christensen et al., 2006). It initially consisted of three identical satellites (Swarm A, Swarm B, and Swarm C) which were launched into Low Earth Orbit in 2013. Initially, the spacecraft flew in a string-of-pearls configuration before the final constellation of the mission was achieved on 17th April 2014. Swarm A and C formed the lower pair of satellites, which flew in close proximity at an altitude of ~462 km, whereas Swarm B was at ~511 km. Despite being mainly conceived as a magnetic mission, Swarm also observes the ionospheric plasma. A large number of papers have been published in this field and these have been reviewed by Wood et al. (2022). The configuration of the Swarm satellites, their near-polar orbits and the data products developed, enable studies of the spatial variability of the ionosphere at multiple scale sizes (Kotova et al., 2022). A range of data products to characterise this variability were developed from the Swarm observations as part of the project "Ionospheric Plasma Irregularities Characterized by the Swarm Satellites – IPIR". IPIR combines data from different instruments on board the Swarm satellites, which act as proxies for the plasma density variations in the ionosphere along the satellite's trajectories at multiple scale sizes (Jin et al., 2019, 2022). Multiscale analysis was used to determine the dominant scales of the plasma structures when observed at each of these scale sizes (Urbar et al., 2022). One of the IPIR products is the IPIR index (IPIR_ix), a categorical variable based upon both the rate of change and the standard deviation of the electron density. The IPIR index can also be an indicator of plasma variations, which can lead to scintillation effects. This was demonstrated by Kotova et al. (2023), by comparing data from 23 ground-based scintillation receivers at polar, auroral and low latitudes with data from the Swarm satellites. While these products are not produced fast enough to provide operational nowcasting at present, they do lay the foundations for such operational services in the future (Jin et al., 2020).

The purpose of this paper is to describe the development of a series of statistical models, which predict the variability of ionospheric plasma. Such models are designed to advance the physical understanding of the system and to lay the foundations for an operational tool, which can infer the behaviour of the ionosphere in regions scarcely covered by ground-based instrumentation. Additionally, as corroborated by the statistical work of Kotova et al. (2022), modelling of the plasma quantities available in the IPIR product can support GNSS-based studies of ionospheric irregularities and their effect on L-band signals. Two versions of the models are produced. The first version is based solely upon data products which are available in either real-time or near real-time, to move towards an operational model and assess the performance of such a model. The second version of the models includes other observations which are not so readily available, to determine what product(s) may be useful to develop for future operational services.

The paper is structured as follows: Section 1 gives an overview of the background literature, Section 2 describes the development of the models and Section 3 describes the process of model optimisation and evaluation. The results are discussed in Section 4 and conclusions are drawn in Section 5. The companion paper, Spogli et al. (2024), which is hereafter referred to as Paper 2, assesses the performance of the models created within the present paper.

2 Model development

2.1 Overview of method

The technique of Generalised Linear Modelling (GLM) (McCullagh & Nelder, 1983) has been applied in numerous fields including medical trials (e.g. Schwemer, 2000), road safety (e.g. Wood et al., 2013) and ionospheric physics (e.g. Dorrian et al., 2019). A special case of a GLM is a linear model, whereby a dependent variable is predicted from an explanatory variable using an equation of the form:

$$E(y) = \beta_0 + \beta_1 \cdot x_1, \tag{1}$$

$E(y)$ is the expected value of dependent variable y, which is to be predicted, x_1 is the explanatory variable and β_0 and β_1 are empirically determined constants known as the parameter estimates. It is postulated that the explanatory variable influences the dependent variable, and so the dependent variable can be predicted from the explanatory variable. Many systems have dependent variables which are influenced by multiple explanatory variables and multivariate linear models, which are another special case of a GLM, are commonly used in such cases. In such models, the dependent variable is predicted from several explanatory variables, using an equation of the form:

$$E(y) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n, \tag{2}$$

$x_1 \ldots x_n$ are the explanatory variables and $\beta_1 \ldots \beta_n$ are the associated parameter estimates. A GLM is similar to that stated for a multivariate linear model. The differences are that the

Table 1. The dependent variables selected to represent the plasma density and the variability of this plasma. These were all taken from the Swarm level 2 data product IPDxIRR_2F.

Dependent variable	Description	Units
|Grad_Ne@100km|	The electron density gradient in a running window calculated via linear regression over 27 data points for the 2 Hz electron density data.	$cm^{-3}\ m^{-1}$
|Grad_Ne@50km|	The electron density gradient in a running window calculated via linear regression over 13 data points for the 2 Hz electron density data.	$cm^{-3}\ m^{-1}$
|Grad_Ne@20km|	The electron density gradient in a running window calculated via linear regression over 5 data points for the 2 Hz electron density data.	$cm^{-3}\ m^{-1}$
IPIR_ix	The product of the rate of change density index in 10 s (RODI10s) and the standard deviation of the electron density in a running window of 10 s ($A(n_e)_{10s}$).	None
Ne	Electron density.	cm^{-3}

dependent variable is not assumed to follow a normal (Gaussian) distribution and that the link function (the form of the equation) may also change. It is commonly expressed as:

$$g(E(y)) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n, \quad (3)$$

where $g(E(y))$ is a function of the expected value of the dependent variable. In the present paper, GLMs were used to create a series of statistical models of the ionospheric plasma and measures of the variability of this plasma.

2.2 Choice of dependent variables

A number of dependent variables were chosen, as shown in Table 1. |Grad_Ne@100km|, |Grad_Ne@50km| and |Grad_Ne@20km| were selected as these act as proxies for the variability of ionospheric plasma at spatial scales of 100 km, 50 km and 20 km respectively. These were taken from the Swarm level 2 data product IPDxIRR_2F (Jin et al., 2022), which is available at: ftp://swarm-diss.eo.esa.int. The absolute value of these values was used to ensure that this measure was not dependent upon the direction in which the satellite was moving. The IPIR index, which is a categorisation of fluctuations in the ionospheric plasma density (0–3 low, 4–5 medium, and >6 high level), was also selected. This is the product of the rate of change density index in 10 s (RODI10s) and the standard deviation of the electron density in a running window of 10 s ($A(n_e)_{10s}$). Based on the motion of the satellite, this corresponds to a horizontal spatial scale of approximately 80 km. Finally, the plasma density was also selected. This was also taken from the Swarm level 2 data product IPDxIRR_2F, where the electron density was directly copied from the Langmuir probe files and downsampled to 1 Hz to match the data rate of other data products which are available in IPDxIRR_2F. The use of the electron density from IPDxIRR_2F also ensured that all the dependent variables used in the Swarm-VIP project were calculated from the same baseline (baseline 3). It should be noted that, although these data are labelled as electron density within the Swarm data products, it is actually the ion current that is measured for this product as this is the cleaner, more reliable measurement (Buchert, personal communication). The ion density is estimated using Langmuir's orbital-motion-limited (OML) model (Mott-Smith & Langmuir, 1926) with the assumption of O^+ being the dominant ion. The plasma is assumed to be quasi- neutral, and the ion density is currently used as a proxy for the electron density in the Swarm level 1B and level 2 data products (Buchert, personal communication). In the remainder of the paper, as global neutrality of the ionospheric plasma is assumed, the plasma density is referred to as the electron density.

2.2 Choice of explanatory variables

A number of explanatory variables were chosen, and these acted as proxies for the driving processes. For example, a commonly used proxy for solar activity is the F10.7cm solar radio flux, and this was used as a proxy for solar activity. The full list of explanatory variables trialled is given in Table S1 in the Supplementary Material. In essence, these fall into several broad categories:

- *Solar activity*: F10.7cm solar radio flux (observed) and the sunspot number R.
- *Solar wind*: Bulk speed, density, pressure, Interplanetary Magnetic Field (IMF) and Interplanetary Electric Field (IEF).
- *Geomagnetic activity*: The aa, AE, am, AL, Ap, ASY-D, ASY-H, AU, Dst, Kp, Polar Cap (North) index (PCN), SYM-D and SYM-H indices.
- *Location*: Geographic latitude (LAT), magnetic latitude (MLAT), local solar time (ST) and magnetic local time (MLT).
- *Complementary observations from Swarm*: The thermospheric density and current systems.
- *Miscellaneous*: Solar zenith angle (SZA), a function based on the ST to represent the diurnal variation and a function based on day of year (DOY) to represent the seasonal variation.

Two versions of the models for each dependent variable were produced. The first version was based solely upon data products which are available in either real-time or near real-time, to move towards an operational model and assess the performance of such a model. The second version of the models included other observations which are not so readily available, to give a deeper understanding of the physical system and to determine which product(s) may be useful to develop for future operational services. The complete list of which explanatory variables were trialled in which version of the models is given in Table S1 in the Supplementary Material. Many of these are taken from, or calculated from, the OMNI dataset (https://spdf.gsfc.nasa.gov/pub/data/omni/). These included the clock angle and a number of solar wind coupling functions, which are summarised in

Newell et al. (2007). The clock angle, θ_c, shows the relative importance of the y- and z-components of the IMF and is defined as:

$$\theta_c = \arctan\frac{|B_y|}{B_z}. \quad (4)$$

A clock angle of 0° is purely IMF B_z positive with a B_y component of zero, 180° is purely IMF B_z negative with a B_y component of zero and 90° is completely dominated by $|B_y|$ with a B_z of zero. Three solar wind coupling functions were trialled. The first of these was introduced by Newell et al. (2007) and was given by:

$$E_N = v^{4/3} \cdot B_T^{2/3} \cdot \sin^{8/3}\left(\frac{\theta_c}{2}\right), \quad (5)$$

where E_N is the solar wind coupling function, v is the solar wind velocity and B_T is the magnitude of the IMF. The second of these was Akasofu's ε parameter (Akasofu, 1996). This is proportional to:

$$\varepsilon \propto vB_T{}^2 \sin^4\left(\frac{\theta_C}{2}\right). \quad (6)$$

This can also be expressed as $\varepsilon = vB_T{}^2 \sin^4\left(\frac{\theta_C}{2}\right) l_0^2$ where l_0 is an empirically determined scale factor with units of length (Koskinen & Tanskanen, 2002). In the present study, it is an association between ε and the dependent variable which is of interest. The numerical value of ε is irrelevant and the scale factor l_0 has not been used. The third and final of the solar wind coupling functions, E_{LYA}, resulted from a student summer project (Daniel Elliot, personal communication) where the powers in equation (6) were varied and the version which had the most significant statistical relationship to the measure of the variability of polar cap plasma defined by Wood & Pryse (2010) was selected. E_{LYA} was given by:

$$E_{\mathrm{LYA}} = vB_T{}^{\frac{1}{2}} \sin^2\left(\frac{\theta_C}{2}\right). \quad (7)$$

The version of the F10.7cm solar radio flux (Tapping, 2013) present within the OMNI dataset is the adjusted version, which is corrected for variations in the Sun-Earth distance. As the present study is concerned with ionospheric plasma, the flux incident on the Earth is the value of primary interest. Therefore, the observed version was used (data are available at: https://lasp.colorado.edu/lisird/). Also trialled as explanatory variables were the LAT, the MLAT, the ST, the MLT, the SZA and a sine function based on the DOY, going from −1 at midwinter to +1 at midsummer in the northern hemisphere. The purpose of this sine function was to act as a proxy for the annual anomaly.

In the model development, no measure of longitude (geographic or geomagnetic) was trialled as an explanatory variable due to the characteristics of the Swarm orbit. During a year, Swarm samples all local time and longitude sectors. However, it only samples a given local time sector in a given longitude sector once every 131 days, which corresponds to two or three intervals per year. It is not feasible to trial both local time and longitude using a dataset that spans 2 years and, at the time of writing, it was not currently feasible to extend this dataset without compromising the ability of the model to consider times of higher solar activity. However, as the Swarm mission continues during solar cycle 25, then it will be possible to extend the dataset and to trial both longitude and local time as explanatory variables.

As well as observing the ionospheric plasma, the Swarm mission can infer the thermospheric density, the magnitude of the field-aligned currents and the magnitude of the radial currents. These were trialled as explanatory variables within the second version of the models. As the geomagnetic indices AE, AL and AU were only available in the OMNI dataset until 28th February 2018, these were also only trialled within the second version of the models. Two additional geomagnetic indices, *aa* and *am*, which describe the mid-latitude ionosphere were also trialled in the second version of the models.

2.3 Dataset

Two years of data were used for model development, covering 16th July 2014–15th July 2015 and 1st January 2017–31st December 2017. The first of these intervals covered a time of higher solar activity, while the second interval covered a time of lower solar activity. The first interval began on the first date at which the IPDxIRR 2F data product was publicly available at ftp://swarm-diss.eo.esa.int. Whole years of data were used to ensure that all local times and longitude sectors were sampled. The dataset was restricted to 2 years to avoid the times of higher solar activity being under-represented in the dataset. This would have resulted in a reduction of the statistical significance of the relationship between proxies for solar activity and the dependent variable, potentially removing information about this driver from the models.

It was postulated that different driving processes may dominate in different latitudinal regions. Therefore, the dataset was broken into four subsets, to represent the polar, auroral, mid-latitude and equatorial regions respectively. Data were assigned to the appropriate region using the ionospheric region flag in the IPDxIRR 2F data product. The methodology used to determine the ionospheric region was described by Jin et al. (2022). A small amount of data could be misclassified based on the ionospheric region flag alone. Therefore, data were excluded from a particular region if the modulus of the magnetic latitude was outside of the following limits:

- Polar latitudes: 50°–90° MLAT.
- Auroral latitudes: 50°–90° MLAT.
- Mid latitudes: 30°–70° MLAT.
- Equatorial latitudes: 0°–40° MLAT.

The points in the dataset from which the models were developed need to be independent. To ensure the independence of data points, the largest spatial scales commonly observed in |Grad_Ne@100km| were identified. Thirty three days were selected, to cover a range of seasons, geomagnetic activities and local time sectors. All orbits on each day were inspected and the largest plasma structures, defined as the distance between successive times when the conditions Grad_Ne@100km = 0 and Grad_Ne@100km ≥ 0 occurred simultaneously, i.e. when Grad_Ne@100km was zero but also increasing, were identified. This analysis was conducted in four different regions (polar, auroral, mid-latitude and equatorial), with the observations split into each region using the ionosphere region flag in the IPDxIRR 2F data product.

At polar, auroral and mid-latitudes, the largest intervals corresponding to this definition of plasma structure were 142 s, 117 s and 297 s respectively. The latter two of these were rounded up to give intervals of 142 s, 120 s and 300 s respectively. This did not mean that plasma structures of these sizes routinely occur in the ionosphere (the time interval of 300 s in the mid-latitude region corresponds to some 20° of latitude), merely that using these intervals gave confidence that the data are independent. The equatorial region was dominated by the EIA, which spans these latitudes (Rishbeth 1971). Data points within this region are very different from one another. However, based on the criteria by which the independence of |Grad_ Ne@100km| was assessed, they are not independent of one another. A time interval of 75 s (roughly corresponding to 5° of latitude) was selected for this region.

In order to create the database for the polar region, the first 142 s of data in this region during each day were taken and a point was randomly selected for inclusion in the database. Points every 142 s from this point were then selected. The same method (with different time intervals) was used in the other regions.

The databases in the polar, auroral, mid-latitude and equatorial regions comprised 34,404, 65,358, 78,097 and 116,519 points respectively. Datasets for model optimisation and evaluation were also created, using data which was not included in the training dataset. Data from the following dates were used for these datasets:

- 1st January 2014–15th July 2014: Optimisation and evaluation.
- 16th July 2014–15th July 2015: Training.
- 16th July 2015–31st December 2016: Optimisation and evaluation.
- 1st January 2017–31st December 2017: Training.
- 1st January 2018–28th February 2018: Optimisation and evaluation.

Within this optimisation and evaluation dataset, dates where the DOY was an even number were used for optimisation and dates where the DOY was an odd number were used for evaluation. It was intended that each of the optimisation and evaluation datasets would contain one calendar year of data, to cover all seasons, local times and longitude sectors. Data gaps in some of the Swarm data products in early 2014 resulted in the decision to include an additional 2 months of data from early 2018 in these datasets. The final constellation of the mission for science operations was achieved on 17th April 2014. The decision to include data from before this date in the optimisation and evaluation datasets ensured that times of higher solar activity were well represented in these databases. However, as these data were from higher altitudes than those within the training database, this will worsen the model performance. Therefore, the "true" model performance at the altitude of Swarm A is likely to be slightly better than stated in the statistics reported in this paper.

2.4 Choice of distribution for the dependent variables

An appropriate distribution needed to be chosen to represent the dependent variable. Those commonly used to represent continuous data in GLM are the Gaussian (normal), Gamma, lognormal and inverse Gaussian distributions. However, in this study, a greater range of distributions were trialled. These were the Birnbaum Saunders, Burr, Exponential, Extreme Value, Gamma, Inverse Gaussian, Logistic, Loglogistic, Lognormal, Nakagami, Normal, Rician, tLocationScale and Weibull distributions. These distributions were trialled for the dependent variables shown in Table 1, and the ability of these distributions to represent the dependent variable was evaluated by visual inspection of quantile-quantile (QQ) plots. A QQ plot shows the quantiles of the data on the y-axis and the quantiles of the modelled values on the x-axis. If, for example, a normal distribution was trialled, then a mean and standard deviation would be estimated from the data. A distribution of points would then be estimated from the mean and the standard deviation, and the quantiles of these values would be shown on the x-axis. Ideally, the points should be on the $x = y$ line.

None of the distributions trialled adequately represented the data. The example shown in Figure 1 is for |Grad_Ne@100km| in the polar region. For all distributions in all latitudinal regions, the trend shown by the points deviated substantially from the $x = y$ line. In the case of the Gamma distribution (right-hand panel), the higher values of the observations are consistently greater than the model. This suggests that the model will struggle to predict the observations associated with the largest values. Therefore, instead of modelling the dependent variable, the data were transformed to model a function of the dependent variable. Logarithms (natural, base 2 and base 10), e^x, 2^x, 10^x, the nth power and (up to $n = 5$), the nth root (up to $n = 9$) were all trialled, and the resulting QQ plots were manually inspected. The purpose of this exercise was to find a good distribution to represent the dependent variable. It was more important to ensure some measure of consistency between the models than to obtain the very best possible choice of distribution in every case. Inspection of the QQ plots, of which examples are shown in Figures 1 and 2, led to the choice of the n^{th} root. The gamma distribution was used for models of |Grad_Ne@100km|, |Grad_Ne@50km| and |Grad_Ne@20km|. The normal distribution was used for models of electron density. IPIR_ix is a categorical variable taking discrete values, so this was modelled assuming a Poisson distribution. The transformations and distributions chosen are shown in Table 2.

2.5 Choice of link function

There are three link functions which are commonly used with the Gamma distribution. These are the identity link function:

$$E(y) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n, \tag{8}$$

the inverse link function:

$$E(y) = \frac{1}{\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n}, \tag{9}$$

and the log link function:

$$E(y) = \exp\left(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n\right). \tag{10}$$

In order to establish which to use for the dependent variables which were represented by the Gamma distribution, the statistical significance of the relationship between the dependent variable and each explanatory variable was tested for each link

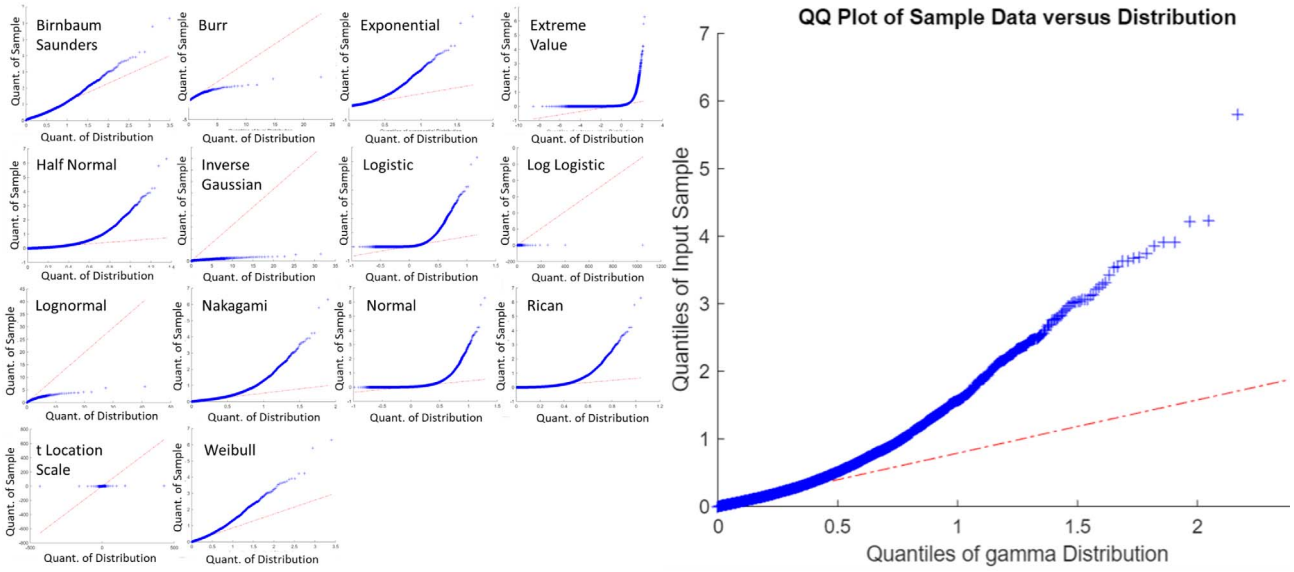

Figure 1. Quantile-quantile (QQ) plots for |Grad_Ne@100km| in the polar region when different distributions are trialled to represent these data. The distributions trialled were: First row, left to right: Birnbaum Saunders, Burr, Exponential and Extreme Value. Second row, left to right: Half normal, Inverse Gaussian, Logistic and Loglogistic. Third row, left to right: Lognormal, Nakagami, Normal and Rician. Fourth row, left to right: tLocation Scale and Weibull distributions. Right-hand panel: Gamma.

function in each latitude range (polar, auroral, mid and equatorial). A score was assigned based on the significance of this relationship:

- If the significance, s, was 0.01% or better, then the score was 4.
- If $0.01\% < s \leq 0.1\%$, then the score was 3.
- If $0.1\% < s \leq 1\%$, then the score was 2.
- If $1\% \leq s < 5\%$, then the score was 1.

For each link function, the average score across all parameters was then found, and the link function with the highest value was selected. On this basis, the log link function was chosen.

The link function commonly used with a normal distribution is the identity link function. In the case of the Poisson distribution, the commonly used choice is the log link function. These were selected for the models of the electron density and IPIR_ix respectively.

2.6 Model fitting procedure

Models were developed for each dependent variable separately. The first step of this process was to fit a single-term GLM for each explanatory variable (i.e. if the database contained n explanatory variables, then n single-term models were fitted). This was conducted using the statistical computing software "R" (version 4.1.1). The glmfit command from the MASS (Modern Applied Statistics with S) package was used. The statistical significance of the relationship between the explanatory variable and the dependent variable was established in each case. The explanatory variable with the most statistically significant relationship to the dependent variable was chosen ("explanatory variable 1"). The statistic used to assess the statistical significance of this relationship was the p-value. If models using different explanatory variables had the same p-value, and if this was the lowest p-value for the explanatory variables tested, then a secondary criterion was needed to choose between this subset of explanatory variables. The secondary criterion was the highest correlation between the dependent variable and the explanatory variable. The explanatory variable chosen was added to the main (overall) model for the dependent variable considered. This model, containing explanatory variable 1, explained some, but not all, of the variability in the dependent variable.

Two term models were then trialled, using a subset of the remaining explanatory variables. The technique of GLM requires explanatory variables to be independent. Therefore, if the correlation between the explanatory variable trialled and any other explanatory variable in the main (overall) model was greater |0.25|, then this explanatory variable was excluded from this analysis. This does not mean that a correlation of 0.26 was considered to be important, but rather a correlation of 0.25 was not considered to be important. The remaining subset of possible explanatory variables was used to create two-term models. Each of these included the dependent variable, explanatory variable 1 and another explanatory variable, with each possible variable considered in turn. The explanatory variable from the two-term model with the greatest statistical significance (lowest p-value) was added to the main (overall) model for this dependent variable. If models using different explanatory variables had the same p-value, and if this was the lowest p-value for the explanatory variables tested, then a secondary criterion was needed to choose between this subset of models. In this case, the secondary criterion was the lowest correlation between explanatory variable 1 and the explanatory variable trialled. The explanatory variable chosen was added to the main (overall) model for the dependent variable considered. The combination of these two explanatory variables explained some, but not all, of the variability in the dependent variable.

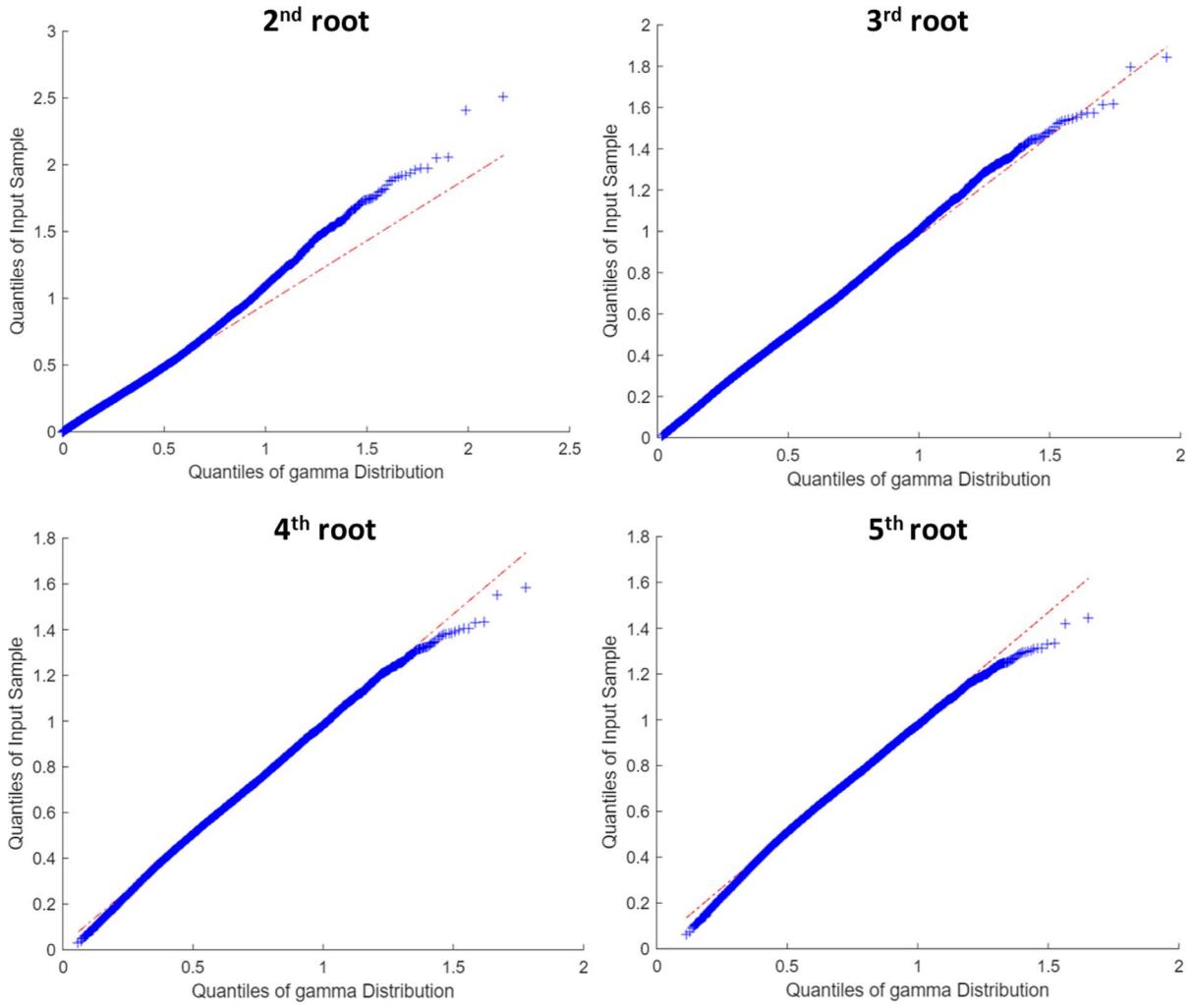

Figure 2. Quantile-quantile (QQ) plots for |Grad_Ne@100km| in the polar region assuming a Gamma distribution when different transformations are trialled to these data. The transformations are 2nd root (upper left panel), 3rd root (upper right panel), 4th root (lower left panel) and 5th root (lower right panel).

Table 2. The transformations applied to the dependent variables used to represent the ionospheric plasma and the variability in this plasma, together with the distributions chosen.

Dependent variable	Distribution	Transformation applied to dependent variable			
		Polar	Auroral	Mid-latitude	Equatorial
|Grad_Ne@100km|	Gamma	3rd root	3rd root	7th root	4th root
|Grad_Ne@50km|	Gamma	3rd root	8th root	7th root	8th root
|Grad_Ne@20km|	Gamma	3rd root	2nd root	2nd root	2nd root
IPIR_ix	Poisson	NA	NA	NA	NA
Ne	Normal	4th root	4th root	4th root	6th root

This process was repeated until no further explanatory variables were statistically significant at the 5% level when added to the model. The model produced shows which combination of the explanatory variables tested best explained the variability in the dependent variable.

2.7 Model optimisation

The models fitted using the process outlined in Section 2.6 contains a large number of terms. As an example, the polar model of |Grad_Ne@100km| was:

$$\sqrt[3]{|\text{GradNe@100km}|} = \exp\left(\beta_0 + \beta_1 \cdot \text{F10.7}_{81} + \beta_2 \cdot \text{SZA} + \beta_3 \cdot \text{fDOY} + \beta_4 \cdot \text{Kp} + \beta_5 \cdot |\text{MLAT}| + \beta_6 \cdot B_x + \beta_7 \cdot \text{SWDen} + \beta_8 \cdot \text{SYMD}\right). \quad (11)$$

An explanation of the terms in the model is given in Table S1 in the Supplementary Material. The process of model optimisation was undertaken to determine whether all of the terms in such equations were justified.

Each model was refitted using the optimisation database. Any terms which were no longer significant at the 5% level or better, were removed. When implementing this method, the least significant term was removed first. The model was then refitted, and the next least significant term was removed if it was not significant at the 5% level. This iterative process continued until the only terms left in the model were significant at the 5% level or better. In this example, namely, the polar model of |Grad_Ne@100km|, two terms (B_x and *SYM_D*) were removed due to this process. One of the dangers of a statistical model is that there is always the possibility of spurious results. When working at the 95% confidence level (5% significance), there is a 5% chance that a result is spurious. The purpose of this first optimisation step is to reduce the chance of spurious results appearing in the models. An explanatory variable must be statistically significant at the 5% level in both the training and optimisation datasets, thus reducing the chance of a spurious term in the model to, at most, 0.25%. This does not guarantee that any terms removed during this process are spurious, it simply means that the statistical relationship between this term and the dependent variable is not strong enough to warrant inclusion in the model. In this example, equation (11) became:

$$\sqrt[3]{|\text{GradNe@100km}|} = \exp\left(\beta_0 + \beta_1 \cdot \text{F10.7}_{81} + \beta_2 \cdot \text{SZA} + \beta_3 \cdot \text{fDOY} + \beta_4 \cdot \text{Kp} + \beta_5 \cdot |\text{MLAT}| + \beta_7 \cdot \text{SWDen}\right). \quad (12)$$

As a next step, Akaike's An Information Criterion (AIC) was used to test the remaining terms (Barlow, 1989). The AIC is a statistic used to evaluate the trade-off between model performance and model complexity. It is calculated from the maximum value of the likelihood function for the model ($\hat{L}$) and the number of fitted parameters (k) and is given by:

$$\text{AIC} = 2 \cdot k - 2 \cdot \ln\left(\hat{L}\right). \quad (13)$$

The optimum solution within a series of nested models is the one with the lowest AIC. For example, if there are (for example) five independent variables in a model, then this can be thought of as five nested models. The first model contains only the first independent variable, the second model contains only the first two independent variables etc.

The AIC is commonly used to determine whether additional complexity in the models is justified, but this is a tool which needs to be carefully interpreted. There are several decades of research work showing that the variability of ionospheric plasma is influenced by solar activity, geomagnetic activity/solar wind, latitude and local time. Therefore, a limit was imposed on what terms could be removed based on the AIC, to ensure that each of these drivers were represented (provided that they were statistically significant). Terms were tested and removed, starting with the nested model with the largest number of terms. This process was stopped when the removal of the term considered would completely remove a key driver (i.e. if the process would remove all proxies for any of the following: Solar activity, geomagnetic activity/solar wind, latitude or local time). In the example of the polar model of |Grad_Ne@100km|, another term was removed as a result of this process. In this case, the complexity added to the model by including *SW_Den* was not justified based on the model performance and equation (12) became:

$$\sqrt[3]{|\text{Grad}_\text{N}\text{e@100km}|} = \exp\left(\beta_0 + \beta_1 \cdot \text{F10.7}_{81} + \beta_2 \cdot \text{SZA} + \beta_3 \cdot \text{fDOY} + \beta_4 \cdot \text{Kp} + \beta_5 \cdot |\text{MLAT}|\right). \quad (14)$$

This process was undertaken for all the models fitted.

2.8 Models created

The models created are summarised in Tables S2, S3 and S4 of the Supplementary Material. Table S2 shows version 1 of the models, based upon independent variables which are available in near real-time. In Tables S2 and S3, two versions of the equatorial models are shown. Version 1 (Table S2) underwent the process of optimisation and evaluation using a subset of the data points within the optimisation and evaluation database. Version 1 of the equatorial models used 116,519 data points, which was greater than the number of data points in any other latitudinal region. All solar, local time and geomagnetic conditions were sampled. After this product was created, further model development activities were undertaken, one of which involved splitting the equatorial database by local time. To maintain a large data volume for optimisation and evaluation, this process used all available data from the years considered. In the interests of completeness, the equatorial model was revised using this larger database, as shown in Table S3. This made relatively little difference to the choice of model terms, their parameter estimates and the model performance.

During the process of assessing the performance of the models in reproducing the known climatological features of the topside ionosphere (reported in Paper 2), it was shown that the equatorial models did not adequately represent EPBs. It is possible that these were not well represented as the model was dominated by variations between day and night. Therefore, it was decided to create three additional categories of model in the equatorial region, one to represent daytime, one to represent nighttime and one to represent the evening when EPBs were more likely to be present. Plots showing the mean, median and standard deviation of |Grad_Ne@100km|, |Grad_Ne@50km|, |Grad_Ne@20km| and the electron density in one-hour blocks were produced (Fig. 3). Inspection of these plots suggested that the three different local time sectors could be set separately as 01-08 LT (night), 08-18 LT (day) and 18-01 LT (bubbles). Table S3 in the Supplementary Material shows the resulting models, with the "all day" equatorial model included for reference. In each case, an appropriate transformation of the data was selected using the method outlined in Section 2.4. The transformation selected is also shown in Table S3 in the Supplementary Material.

Table S4 in the Supplementary Material shows version 2 of the models, which includes additional explanatory variables.

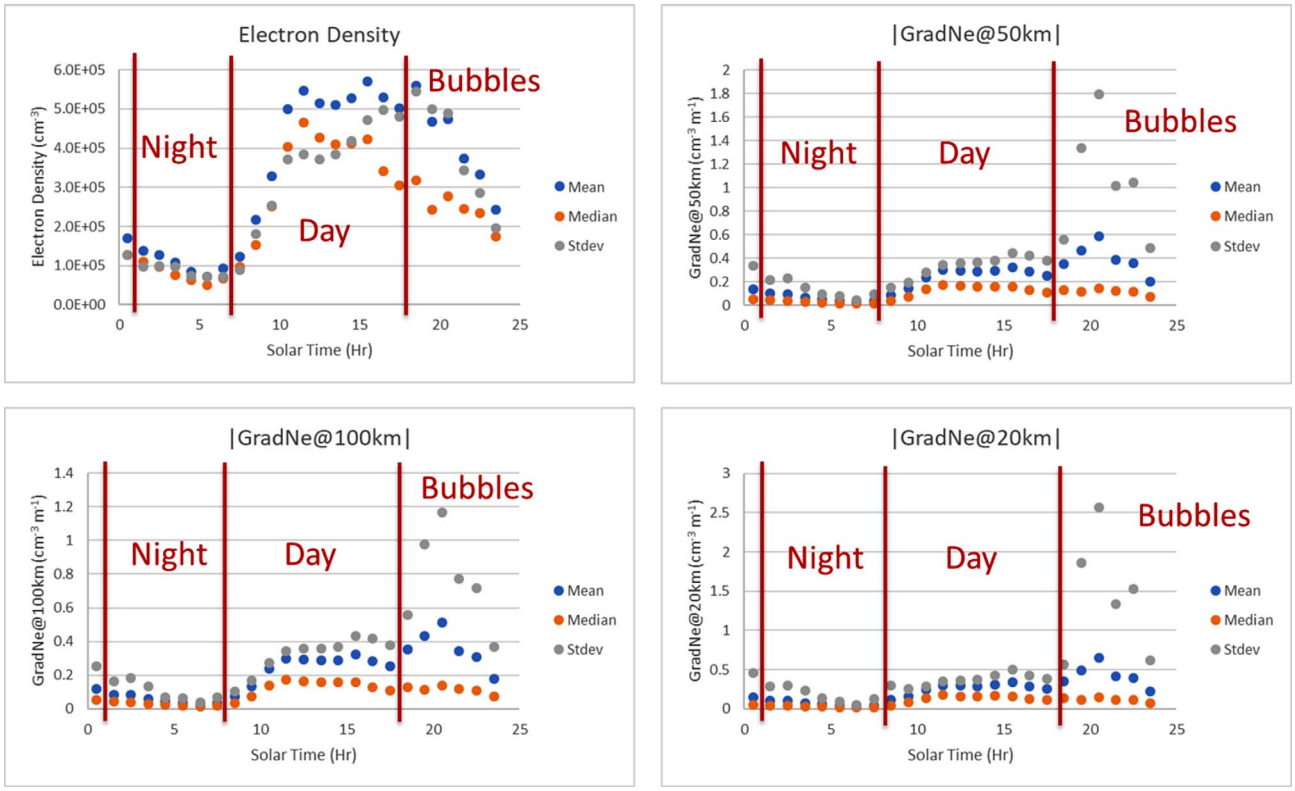

Figure 3. The mean, median and standard deviation of |Grad_Ne@100km|, |Grad_Ne@50km|, |Grad_Ne@20km| and the electron density in one-hour blocks in the equatorial region.

The primary purpose of this second version of the models was to investigate how the inclusion of the thermospheric density affected the model performance. The thermospheric density was determined by precise orbit determination (POD; van den IJssel et al., 2020). If the thermospheric density was not included within a model, then no new model is presented here. In two cases, both within the mid-latitude region, no new model is presented, as the thermospheric density observed by Swarm was correlated with an explanatory variable which became the first term in the model. In the case of the model of |Grad_Ne@20km|, the first term in the model was the F10.7cm solar flux, which had a correlation of 0.73 with the thermospheric density. In the case of the model of IPIR_ix, the first term in the model was the MLT, which had a correlation of 0.26 with the thermospheric density. In both cases, this led to the exclusion of the thermospheric density from the model.

Version 2 of the models also trialled a greater range of explanatory variables than the thermospheric density alone, as summarised in Table S1 in the Supplementary Material. The same model fitting procedure that was used for version 1 of the models was applied. The only additional explanatory variables that became part of version 2 of the models were the thermospheric density, the field-aligned currents (FAC) and the ionospheric radial currents (IRC), which are available as Swarm data products. FAC and IRC only appeared in two models; those of the electron density in the polar and equatorial regions. To allow a clear discussion of the impact of adding the thermospheric density as an explanatory variable, these two models were also re-created without considering FAC and IRC as explanatory variables. The overall purpose of this paper is to build a model capable of reproducing the ionospheric variability at all places and in all geospace conditions, which can potentially be used for operations and nowcasting. Such a model needs to be based on readily available proxies for the physical processes, such as those contained in the OMNI dataset. The purpose of version 2 of the models is to provide a deeper understanding of the underlying physical processes and to identify missing variabilities that affect the model performance.

3 Model evaluation

The models were used to predict the data observed in the evaluation database. A comparison between the predictions and the observations using several goodness-of-fit statistics was used to determine the model performance. However, prior to discussing these statistics, it was useful to examine plots of a subset of the data to illustrate the strengths and limitations of the models. Figure 4 shows a statistical comparison between observations and predictions from the Swarm-VIP models in the 0°–15° longitude sector. This sector was chosen as it covers the European region at mid-latitudes, which is one of the regions used for assessing the model performance in Paper 2. Figure 4 shows comparisons of average values in bins spanning 5° of latitude for |Grad_Ne@100km|, |Grad_Ne@50km|,

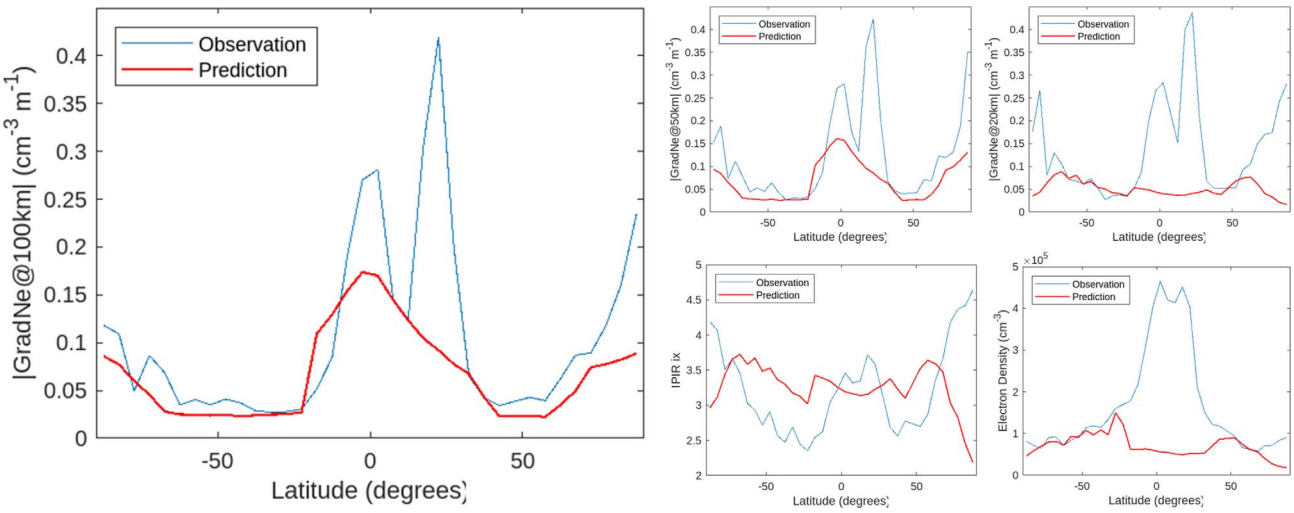

Figure 4. A statistical comparison between observations and predictions from the Swarm-VIP models in the 0°–15° geographic longitude sector for |Grad_Ne@100km| (left panel), |Grad_Ne@50km| (upper middle panel), |Grad_Ne@20km| (upper right panel), the IPIR index (lower middle panel) and the electron density (lower right panel) as a function of latitude. This comparison shows average values in bins spanning 5° of latitude. Negative values of latitude indicate the southern hemisphere. Observations are indicated by the blue lines and predictions by the red lines.

|Grad_Ne@20km|, the IPIR index and the electron density. Observations are indicated by the blue lines and predictions by the red lines. Most points within the evaluation database were used for this comparison, although 339 data points were excluded due to missing data for one or more of the explanatory variables, which prevented predictions from being made. This left exactly 3000 data points which were used for this comparison.

It is immediately apparent from Figure 4 that there are regions of agreement and regions of disagreement between the observations and the model predictions. The comparison for |Grad_Ne@100km| shows that the model captures the variations of this variable at high and mid-latitudes, and also one crest of the EIA. The other crest of the EIA is not captured. A similar pattern is observed for |Grad_Ne@50km|. Models of |Grad_Ne@20km| and the electron density capture the lower values of these variables but not the higher values, particularly in the equatorial region. The IPIR index shows similarities between the predictions and the observations at equatorial latitudes and disagreements elsewhere.

One of the dangers of a statistical comparison of average values of the form shown in Figure 4 is that it can average out regions where substantial variations occur in either the observations or model predictions. In essence, the averages may match but the ranges may not. As an illustration, a comparison of observations and model predictions for a half orbit of the Swarm A satellite was made between 08:09 UT and 08:56 UT on 16th July 2015. This interval was chosen as it is the first half orbit contained within the evaluation dataset, which also sampled the 0°–15° longitude sector and for which the IPDxIRR 2F data product was publicly available at ftp://swarm-diss.eo.esa.int. The start and stop times were determined by the highest latitudes in geographic coordinates. The average geographic longitude of this half orbit was 3.42°. The satellite was moving northwards during this interval. Observations and predictions are presented at a temporal resolution of 1 s, to match the temporal resolution of the IPDxIRR 2F data product. These are shown in Figure 5 while the trajectory of the Swarm A orbit, together with the regions sampled according to the ionospheric region flag in the IPDxIRR 2F data product, are shown in Figure 6. Figure 5, shows that models capture some, but not all, of the trends present in the observations. In all cases, there are observed values that exceed those predicted. A series of sharp discontinuities are present in the model predictions, corresponding to the boundaries between different regions of the ionosphere, as identified by the ionosphere region flag in the IPDxIRR 2F data product. In the southern hemisphere, these boundaries are located at 79.8° S (auroral–mid-latitude boundary) and 19.5° S (mid-latitude-equatorial boundary). In the northern hemisphere, the boundaries are located at 36.7° N (equatorial-mid-latitude boundary), 76.5° N (mid-latitude-auroral boundary) and 82.3° N (auroral-polar boundary). The polar region in the southern hemisphere is not sampled within this half orbit as the boundary between the polar and auroral regions was located at 73.5° MLAT, which, in geographic coordinates, was in the previous half orbit, as illustrated in Figure 6. Figure 5 illustrates some of the successes and limitations of statistical models. The model predictions vary at the same rate as the variations of the explanatory variables, which are used as proxies for the driving conditions. For example, the model of |Grad_Ne@100km| (Eq. (14)) includes SZA and |MLAT| as explanatory variables, which vary slightly between adjacent data points, contributing to capturing a smooth, underlying trend. Another explanatory variable in this model is Kp. This variable has a temporal resolution of 3 h, so a single value of Kp = 2 is used for all of the predictions in Figure 5. This low-to-moderate value of Kp is associated with variable values of |Grad_Ne@100km| in the polar region. The model can go some way towards capturing the average value of |Grad_ Ne@100km| in this region but cannot capture the variability due to the temporal resolution of the relevant explanatory variable (Kp). The other explanatory variables in this model,

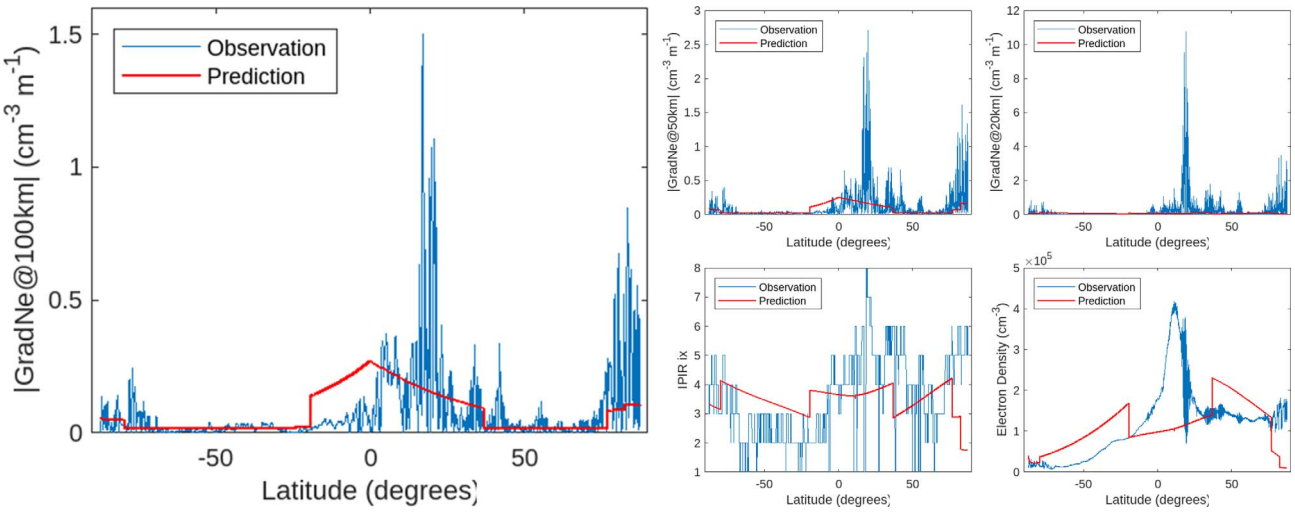

Figure 5. As figure 4, but for a half orbit of the Swarm A satellite between 08:09 UT and 08:56 UT on 16th July 2015. Observations and predictions are presented at a temporal resolution of 1 second, to match the temporal resolution of the IPIR data product. The average geographic longitude of this half orbit was 3.42°.

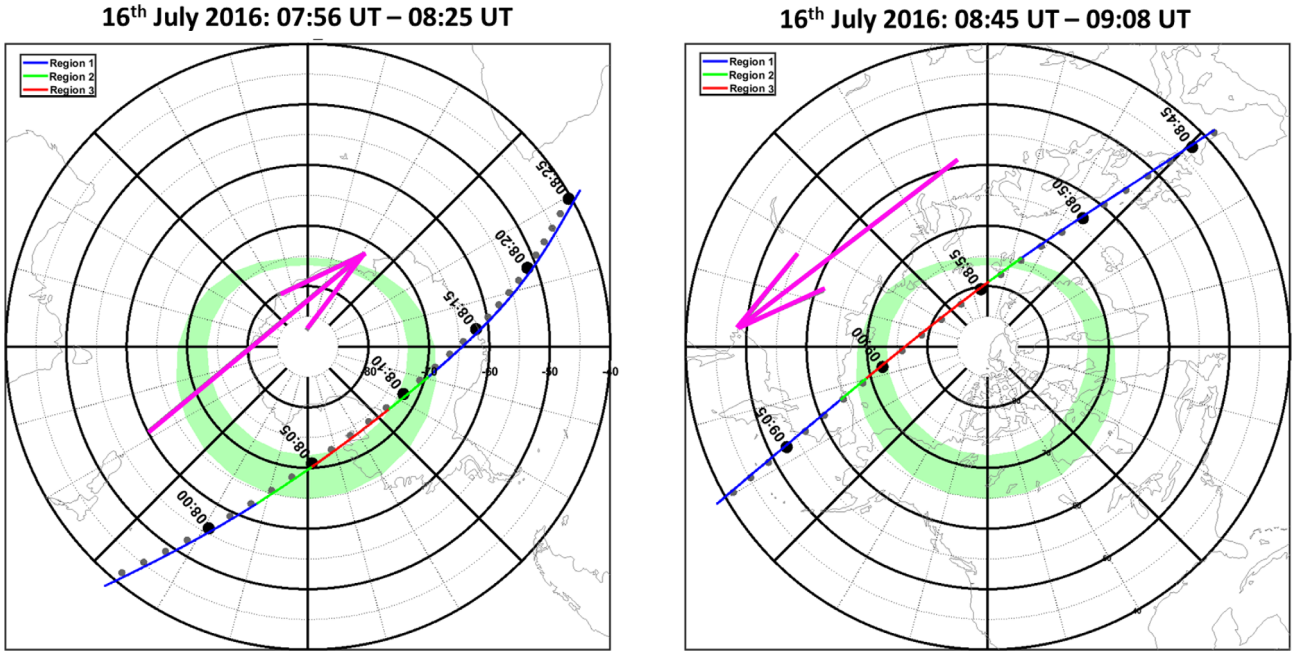

Figure 6. The trajectory of the Swarm A orbit on 16th July 2015 between 07:56 UT–08:25 UT (left-hand panel) and 08:45 UT–09:08 UT (right-hand panel). The plots are centred on the geomagnetic south pole (left-hand panel) and the geomagnetic north pole (right-hand panel). The direction of the satellite motion is shown by the pink arrow. The satellite tracks are colour coded based on the ionospheric region flag in the IPDxIRR 2F data product, with blue representing region 1 (mid-latitude), green representing region 2 (auroral latitudes) and red representing region 3 (polar latitudes).

$F10.7_{81}$ and fDOY, take one value for this day, so they influence the average value of the model prediction shown in Figure 5, but not the short-term variations present in the observations. A more detailed discussion of model performance and the drivers is given in Paper 2, however, it is clear that model evaluation needs to be based on a range of goodness-of-fit statistics. These statistics need to compare not just the average values of the observations and model predictions, but also evaluate whether the models can capture the trends and ranges of values present within the observations.

Liemohn et al. (2021) have discussed goodness-of-fit statistics and their application to statistical models in detail. Four key

measures of the goodness-of-fit of the model predictions to the data were used in the present study.

3.1 Accuracy

This is a measure of the closeness of the model predictions to the observed values. The measures of the accuracy selected in the present study are the relative Root Mean Square Error (rRMSE, the RMSE divided by the median of the observed values) and the Median Symmetric Accuracy (MSA). The RMSE and the MSA are given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N}(M_i - O_i)^2}. \tag{15}$$

$$\mathrm{MSA} = 100 \cdot \left(\exp\left[\mathrm{Median}\left(\left|\ln\left(\frac{M_i}{O_i}\right)\right| \forall i\right)\right] - 1\right). \tag{16}$$

Model values are denoted by M, with individual values with the number set listed as M_i. Observational values are given the variable O, with individual data points called out by O_i. The total number of pairs in the data-model set is N and d is the number of degrees of freedom in the model configuration. The RMSE values are not comparable between models of different latitude regions as this statistic scales with the value of the dependent variable, which spans a different range of values at different latitudes. The rRMSE is more useful as this is comparable between the different models. It is also more intuitive; if rRMSE > 1 then the errors are larger than the predictions. However, in both RMSE and rRMSE, larger values of predictions or observations have a disproportionately greater effect on the statistics. A small number of very large outliers can be responsible for the very large values of rRMSE (i.e. model performance can be good everywhere, apart from a few isolated cases, but the rRMSE suggests that the model performance is poor).

The MSA avoids this drawback, by weighting all points equally and expressing them as a percentage. If this is greater than 100%, then the errors are larger than the predictions. The disadvantage of the MSA is that it can hide issues with the model under particular conditions. A small number of very large outliers have almost no effect on the MSA i.e., the model may not represent some extreme conditions well at all, but the MSA could suggest that model performance is good. Therefore, the rRMSE and MSA in combination give a good assessment of the accuracy of a model.

3.2 Bias

This is a measure of whether the model consistently overpredicts or underpredicts the observations. The statistic used to evaluate this in the present study is the Mean Error (ME), which is given by:

$$\mathrm{ME} = \bar{M} - \bar{O}. \tag{17}$$

If the ME is close to zero, then the models are not significantly biased. If it is greater than zero, then the model consistently overpredicts. If it is less than zero, then the model consistently underpredicts. As with the RMSE, the bias is shown on a relative scale to enable comparisons between different models.

3.3 Precision

This compares the spread of the observations and model predictions and is given by the ratio of the standard deviations of the model and observed values:

$$P_{\sigma,\mathrm{ratio}} = \frac{\sigma_M}{\sigma_O}. \tag{18}$$

If the precision is substantially greater than 1, then the spread of the model predictions is larger than expected (it is likely that the model is too noisy). If the precision is substantially less than 1, then the spread of the model values is lower than the spread of the observations (it is likely that the model is overfitted).

3.4 Association

This measures the association of the observations and predictions, i.e. whether the trends in the observations are captured by the model. In this study, the Pearson Linear Correlation Coefficient was used. This is given by:

$$R = \frac{\sum(O_i - \bar{O})\cdot(M_i - \bar{M})}{\sqrt{\sum(O_i - \bar{O})^2 \cdot \sum(M_i - \bar{M})^2}}. \tag{19}$$

This shows what proportion of the trends in the observations are captured by the model on a scale of 0–1, where 0 indicates that none of the trends in the observations are captured by the model and 1 indicates that the trends are perfectly captured.

The goodness-of-fit of the models are shown in Tables 3–5. The statistics for versions 1 and 2 of the models (Tables 3 and 5) can be directly compared with one another and comparisons are drawn in the following section of this paper. The statistics for the equatorial models in the three local time sectors (Table 4) are all evaluated against different datasets (depending on the local time sector considered), so are not comparable. The purpose of the evaluation in Table 4 was to determine whether the local time sector model could capture the variability associated with EPBs, rather than to draw comparisons to the other models themselves. The ability of the models to capture this variability is discussed in detail in Paper 2.

4 Results and discussion

Collectively, the models show the overwhelming importance of a measure of solar activity as an explanatory variable. In version 1 of the models (Table 3), the 81-day average of the F10.7cm solar radio flux is the first term in 13 out of 20 models, with the daily version of this index selected in a further three cases. These results indicate that this proxy for the driving process is the single most effective term in explaining the observed variability. The modelling approach used within this study builds up the model one term at a time and is particularly appropriate for such a situation. The importance of the F10.7cm solar radio flux could be due to the direct effect of variations in photoionisation, or changes to the chemical composition of the atmosphere. A measure of the position of the observation (LAT or MLAT) or the relative position of the Sun and the observation (DOY_fn, ST_fn or SZA) feature as the first or second term in each of the models. It is interesting to note that proxies for the

Table 3. Goodness-of-fit statistics for version 1 of the models. The goodness-of-fit statistics chosen are the root mean square error (RMSE) on a relative scale (rRMSE, RMSE divided by the median of the observed values), the median symmetric accuracy (MSA), the mean error (ME) on a relative scale (rME, ME divided by the median of the observed values), the precision and the correlation.

Model		Goodness of fit				
Region	Dependent variable	rRMSE	MSA	rME	Precision	Correlation
Polar	|Grad_Ne@100km|	0.47	135	0.00	0.37	0.36
	|Grad_Ne@50km|	0.49	136	0.00	0.37	0.35
	|Grad_Ne@20km|	0.58	141	0.00	0.31	0.30
	IPIR_ix	0.24	115	0.00	0.46	0.45
	Ne	0.16	110	0.00	0.76	0.75
Auroral	|Grad_Ne@100km|	0.47	135	0.00	0.31	0.31
	|Grad_Ne@50km|	0.18	112	0.00	0.28	0.28
	|Grad_Ne@20km|	0.98	168	0.01	0.26	0.24
	IPIR_ix	0.24	115	0.00	0.37	0.36
	Ne	0.15	110	0.00	0.76	0.75
Mid-latitude	|Grad_Ne@100km|	0.19	112	0.00	0.24	0.23
	|Grad_Ne@50km|	0.19	113	0.00	0.22	0.21
	|Grad_Ne@20km|	0.98	158	0.00	0.20	0.20
	IPIR_ix	0.30	126	0.00	0.33	0.34
	Ne	0.16	111	0.00	0.73	0.68
Equatorial version A	|Grad_Ne@100km|	0.41	131	0.00	0.32	0.32
	|Grad_Ne@50km|	0.20	114	0.00	0.34	0.30
	|Grad_Ne@20km|	1.15	169	−0.01	0.46	0.26
	IPIR_ix	0.32	122	0.00	0.27	0.26
	Ne	0.15	111	0.00	0.55	0.55
Equatorial version B	|Grad_Ne@100km|	0.43	131	−0.05	0.57	0.30
	|Grad_Ne@50km|	0.21	114	−0.02	0.56	0.29
	|Grad_Ne@20km|	1.16	170	−0.12	0.61	0.28
	IPIR_ix	0.34	123	−0.04	0.48	0.19
	Ne	0.16	111	−0.03	0.83	0.53

Table 4. As Table 3, but for equatorial models in three local time sectors. These are dayside (08-18 LT), bubbles (18-01 LT) and nightside (01-08 LT).

Model		Goodness of fit				
Region	Dependent variable	rRMSE	MSA	rME	Precision	Correlation
Equatorial	|Grad_Ne@100km|	0.43	131	−0.05	0.57	0.30
	|Grad_Ne@50km|	0.21	114	−0.02	0.56	0.29
	|Grad_Ne@20km|	1.16	170	−0.12	0.61	0.28
	IPIR_ix	0.34	123	−0.04	0.48	0.19
	Ne	0.16	111	−0.03	0.83	0.53
Equatorial: Dayside	|Grad_Ne@100km|	2.76	246	−0.01	0.47	0.36
	|Grad_Ne@50km|	2.75	246	−0.01	0.46	0.35
	|Grad_Ne@20km|	0.79	156	−0.05	0.44	0.40
	IPIR_ix	0.27	116	0.00	0.35	0.35
	Ne	0.32	123	0.01	0.72	0.74
Equatorial: Bubbles	|Grad_Ne@100km|	0.36	126	0.00	0.45	0.42
	|Grad_Ne@50km|	0.24	116	0.00	0.42	0.40
	|Grad_Ne@20km|	0.32	121	0.00	0.39	0.37
	IPIR_ix	0.33	118	0.01	0.33	0.33
	Ne	0.25	117	0.00	0.66	0.61
Equatorial: Nightside	|Grad_Ne@100km|	0.23	115	0.00	0.30	0.30
	|Grad_Ne@50km|	0.23	115	0.00	0.29	0.29
	|Grad_Ne@20km|	0.41	125	0.00	0.27	0.27
	IPIR_ix	0.29	126	0.00	0.28	0.29
	Ne	0.16	112	0.00	0.42	0.43

Table 5. As Table 3, but for version 2 of the models.

Model			Goodness of fit				
Region	Dependent variable	Version	rRMSE	MSA	rME	Precision	Correlation
Polar	|Grad_Ne@100km|		0.50	136	0.00	0.57	0.38
	|Grad_Ne@50km|		0.52	137	0.00	0.44	0.36
	|Grad_Ne@20km|		0.61	143	−0.02	0.38	0.32
	IPIR_ix		0.25	116	−0.01	0.66	0.45
	Ne	With currents	0.72	112	−0.03	2.88	0.16
		Without currents	0.18	111	−0.02	0.70	0.73
Auroral	|Grad_Ne@100km|		0.45	134	0.00	0.48	0.23
	|Grad_Ne@50km|		0.17	112	−0.01	0.45	0.24
	|Grad_Ne@20km|		0.96	168	−0.07	0.49	0.18
	IPIR_ix		0.20	114	−0.03	0.63	0.32
	Ne		0.15	112	0.06	0.74	0.64
Mid-latitude	|Grad_Ne@100km|		0.20	113	0.01	0.20	0.14
	|Grad_Ne@50km|		0.21	114	0.01	0.18	0.13
	|Grad_Ne@20km|				No model fitted		
	IPIR_ix				No model fitted		
	Ne		0.17	112	0.03	0.64	0.71
Equatorial	|Grad_Ne@100km|		0.87	146	−0.49	2.02	0.40
	|Grad_Ne@50km|		0.30	119	−0.17	1.29	0.43
	|Grad_Ne@20km|		2.88	210	−1.28	2.84	0.27
	IPIR_ix		0.35	119	−0.08	0.54	0.28
	Ne	With currents	0.13	111	−0.06	0.85	0.66
		Without currents	0.23	116	−0.16	1.35	0.70

solar wind or geomagnetic activity do not appear in version 1 of the models until term 3 at the earliest, which shows that these proxies are not the dominant variables for explaining the observed variations.

The rRMSE values for all versions of the models fitted (Tables 3–5) are, for the most part, substantially less than 1. This suggests that a reasonable degree of accuracy is obtained by these models. However, the values of the MSA are all greater than 100%, suggesting that the accuracy of the models is poor. This apparent discrepancy can be explained by understanding the differences between rRMSE and MSA. The MSA weights all points equally, while the rRMSE weights larger differences more heavily. The rRMSE suggests that the models represent disturbed conditions reasonably well, providing that they occur reasonably frequently in the dataset. A statistical model of this type cannot capture extreme events that only occur rarely. The large values of the MSA are attributed to substantial percentage differences between predicted and observed values during quiet conditions, but these do not correspond to large absolute differences. The models show relatively little bias. The only model where the bias is substantial is that of |Grad_Ne@20km| in the equatorial region, where the model consistently underpredicts the observations. The precision of most models is substantially less than 1, so the spread of model values is less than the observations. This indicates that the models do not capture the full range of values which are observed. The variations which are not modelled may be due to rarely occurring extreme events or variations driven by a process that is not included in the models. The correlations are substantially less than 1 in most cases, indicating that the trend observed in the data is only partially captured by the models. As the precision indicates that the models do not capture the full range of values observed, likely this is also the reason for the low values of the correlation.

The goodness-of-fit statistics for the equatorial models which are broken into LT sectors show relatively little improvement compared to the equatorial model which covers the entire day. The performance of these models is discussed in detail in Paper 2.

A comparison of the goodness-of-fit statistics between versions 1 and 2 of the models is shown in Table 6. The changes in the measures of accuracy (rRMSE and rME) and correlation were found from simple differences; the changes in the measure of bias (rME) were found by taking the absolute difference compared to zero and the changes in the precision were found by taking the absolute difference relative to one. The purpose of calculating the changes in this way was so that improved model performance in version 2 of the models, which include observations from Swarm, would be indicated by positive values.

In most cases, changes to the accuracy and the bias of the models were small. However, in a number of cases, the use of observations from Swarm as explanatory variables improved other measures of the model performance. In the polar and auroral regions, the addition of the thermospheric density improved the precision of models of ionospheric variability (|Grad_Ne@100km|, |Grad_Ne@50km|, |Grad_Ne@20km| and IPIR_ix). This suggested that, for these models, the addition of the thermospheric density as an explanatory variable led to more of the variability of the system being captured by the models. In the equatorial region, the correlation of almost all of the models improved when observations from Swarm were included, the exception being the model of |Grad_Ne@20km|. This suggested that, in this region, more of the trend in the observations was being captured by the models. The inclusion of current systems in addition to the thermospheric density did not substantially improve the model performance. Current systems were only included in two models, which were the models of electron density in the equatorial and polar regions.

Table 6. Differences in goodness-of-fit statistics between versions 1 and 2 of the models. Positive values indicate larger values of the goodness-of-fit statistics in version 2 of the models.

Model			Goodness of fit				
Region	Dependent variable	Version	ΔrRMSE	ΔMSA	ΔrME	ΔPrecision	ΔCorrelation
Polar	|Grad_Ne@100km|		−0.03	−1.3	0.00	0.19	0.02
	|Grad_Ne@50km|		−0.03	−1.2	0.00	0.07	0.01
	|Grad_Ne@20km|		−0.03	−2.0	−0.02	0.08	0.02
	IPIR_ix		−0.02	−0.8	−0.01	0.20	0.00
	Ne	With currents	−0.57	−2.0	−0.02	−1.64	−0.59
		Without currents	−0.02	−0.7	−0.01	−0.06	−0.02
Auroral	|Grad_Ne@100km|		0.01	0.6	0.00	0.17	−0.08
	|Grad_Ne@50km|		0.00	0.3	−0.01	0.17	−0.04
	|Grad_Ne@20km|		0.02	−0.3	−0.06	0.23	−0.06
	IPIR_ix		0.03	1.1	−0.03	0.26	−0.05
	Ne		0.00	−1.3	−0.06	−0.02	−0.10
Mid-latitude	|Grad_Ne@100km|		−0.01	−0.8	−0.01	−0.03	−0.09
	|Grad_Ne@50km|		−0.01	−0.8	−0.01	−0.04	−0.08
	|Grad_Ne@20km|				No version 2 model fitted		
	IPIR_ix				No version 2 model fitted		
	Ne		−0.01	−0.4	−0.03	−0.09	0.03
Equatorial	|Grad_Ne@100km|		−0.44	−15.5	−0.44	−0.60	0.10
	|Grad_Ne@50km|		−0.09	−4.9	−0.15	0.15	0.14
	|Grad_Ne@20km|		−1.72	−40.1	−1.16	−1.45	−0.01
	IPIR_ix		−0.01	3.4	−0.04	0.06	0.10
	Ne	With currents	0.03	0.3	−0.03	0.02	0.13
		Without currents	−0.06	−5.4	−0.13	−0.18	0.17

In the equatorial region this slightly improved model performance in four out of five of the goodness-of-fit statistics. However, the model performance worsened in the polar region.

While the inclusion of the thermospheric density improved model performance in some cases, there are some substantial limitations in this dataset. The temporal resolution of this dataset is 30 s, which, when the motion of the satellite is considered, corresponds to a spatial resolution of ~2° of latitude. However, the temporal resolution of the densities themselves is ~20 min (van den IJssel et al., 2020), which corresponds to approximately 80° of latitude. The thermospheric density is highly correlated with the F10.7cm solar flux, with correlation coefficients of 0.73, 0.72, 0.69 and 0.65 in the polar, auroral, mid-latitude and equatorial regions, respectively. This indicates that the thermospheric density product used in these models is primarily capturing the large-scale bulk properties of the thermosphere, not smaller-scale structures. Smaller-scale structures in the thermosphere can influence the ionosphere, for example, gravity waves are associated with TIDs (Hunsucker, 1982). There is a Swarm thermospheric density product calculated from non-gravitational accelerations which is available at a higher temporal (and hence spatial) resolution (Bezděk et al., 2018). This is available at a 10-second resolution, corresponding to a horizontal spatial distance of ~80 km, which is similar to the scale sizes of many of the plasma density variations considered in the present paper. This data product may lead to improvements in model performance however, at present, it is only available for Swarm C and contains significant gaps in the usable data. It is hoped to trial this data product as an explanatory variable in a subsequent study. This will require careful and substantial work to ensure that the data gaps do not introduce a selection effect based on local time or latitude into the models and go beyond the scope of the present study.

A perfect fit of the models to the data is neither expected nor observed. These models of the plasma structures are deterministic. However, there are also random variations in the ionospheric structures which cannot be captured by these models. Furthermore, the explanatory variables are proxies for the driving processes. These proxies approximate these processes, rather than exactly replicating them, resulting in a discrepancy. In addition, it could be argued that some of the proxies, such as geomagnetic indices, better represent conditions in the E-layer/around the F-layer peak rather than in the topside ionosphere. Finally, there is no good proxy within the models produced within this paper which could be used for the effect of atmospheric waves and their impact upon ionospheric plasma. Nevertheless, it seems likely that the model performance could be improved by a better specification of the thermosphere.

The statistical models created in this paper test a range of explanatory variables, which are proxies for the driving processes. If a driving process is missing, then this will reduce the performance of the models. In a previous statistical modelling study of the high-latitude ionosphere, Dorrian et al. (2019) showed that the thermospheric temperature was a key term in models which predict the variability of ionospheric plasma. The Gravity Recovery and Climate Experiment (GRACE) and GRACE Follow-on (GRACE-FO; Landerer et al., 2020) mission observes both temperature and winds, which could be tested as explanatory variables within statistical models.

Another limitation of statistical models is that their ability to respond to changes in the driving conditions is determined by the temporal resolution of the explanatory variables which have been used as proxies for the driving processes. For example, the model of |Grad_Ne@100km| (Eq. (14)) included Kp as an explanatory variable. Kp was a better choice than any of the other proxies for geomagnetic activity based on the model

fitting procedure, however, the model cannot respond to changes in the driving process on a timescale of less than the temporal resolution of this variable. As shown in Figure 5, the model can go some way towards capturing the average value of |Grad_Ne@100km| in the polar region but cannot capture the variability. A potentially useful avenue for future research would be to use quantile regression which essentially uses a proxy for the upper boundary of the observed variations as the dependent variable. Quantile regression would allow particular quantiles to be modelled, hence the likely range of the dependent variable to be predicted. The critical discussion of the model's capabilities to reproduce the expected climatological features of the topside ionosphere, in supporting GNSS-based ionospheric observations and its performance against TIE-GCM, is provided in a companion paper (Paper 2).

5 Conclusions

This paper presents a series of statistical models which predict the variability of ionospheric plasma in the topside ionosphere. These models were created by applying the technique of GLM, where measures of the ionospheric plasma and structures within this plasma, were used as the dependent variables. Proxies for the driving processes were used as explanatory variables. Two versions of these models were produced, shown in Tables S2, S3 and S4 in the Supplementary Material. The first version (Tables S2 and S3) is based solely upon data products which are available in either real-time or near real-time, to move towards an operational model and assess the performance of such a model. The first and most significant term in the majority of the models was a proxy for solar activity. The most common second term varied with the latitudinal region. The second term was the SZA in the polar region, a measure of latitude in the auroral region, solar time in the mid-latitude region and a measure of latitude in the equatorial region. Other, less significant terms in the models covered a range of proxies for the solar wind, geomagnetic activity and location. The models are not biased with a mean error of zero to two decimal places in 14 out of 20 cases. The models show a reasonable degree of accuracy with rRMSE as low as 0.15 in particular cases. However, based on measures of the precision and the association, these models do not fully capture the variability present within the observations (Tables 3 and 4).

The second version (Table S4 in the Supplementary Material) of the models includes trialling the thermospheric density and the ionospheric current systems as explanatory variables. The inclusion of the thermospheric density improves the ability of the models to capture the variability observed within the ionosphere in some cases, however, the thermospheric density product only captures the bulk properties of the neutral atmosphere. These models are shown in Table 5. It would be advantageous to use a measure of thermospheric density at a higher temporal, and hence spatial, resolution, and to trial other measures of the thermosphere, such as the temperature and/or velocity. The ability of statistical models to respond to changes in the driving conditions is determined by the temporal resolution of the explanatory variables which have been used as proxies for the driving processes. If the process for which the explanatory variable acts as a proxy results in variability in the dependent variable, then the model can go some way towards capturing the average value of the dependent variable, but not the variability. For example, Kp has a temporal resolution of three hours and it is well known that elevated values of Kp are associated with variability of ionospheric plasma in the polar region. An elevated value of Kp can result in an elevated value of the dependent variable in a statistical model, but create variability in that model on a timescale of less than three hours. A potentially useful avenue for future research would be to use quantile regression to model a proxy for the upper boundary of the likely values.

During a year, Swarm samples all local time and longitude sectors, however, it only samples a given local time sector in a given longitude sector once every 131 days, which corresponds to two or three intervals per year. In the present study, it was not feasible to trial both local time and longitude as explanatory variables within the models without compromising the ability of the model to capture variations in solar activity. The continuation of the Swarm mission into solar cycle 25, makes it possible to extend the dataset and to trial both longitude and local time as explanatory variables and it is anticipated that this will improve the model performance.

Acknowledgements

The Swarm data products are available at ftp://swarm-diss.eo.esa.int, the OMNI resolution data are available at https://spdf.gsfc.nasa.gov/pub/data/omni/, the Service International des Indices Géomagnétiques data are available at https://isgi.unistra.fr/geomagnetic_indices.php and the Laboratory for Atmospheric and Space Physics data are available at https://lasp.colorado.edu. The assistance of Katherine Wood with the proofreading of the manuscript is gratefully acknowledged. The editor thanks Kaleekkal Unnikrishnan and an anonymous reviewer for their assistance in evaluating this paper.

Funding

This work is within the framework of the Swarm Variability of Ionospheric Plasma (Swarm-VIP) project, funded by ESA in the "Swarm+4D-Ionosphere" framework (ESA Contract No. 4000130562/20/I-DT). YJ, DK and WJM acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Consolidator Grant agreement No. 866357, POLAR-4DSpace).

Supplementary material

The supplementary materials of this article are available at https://www.swsc-journal.org/10.1051/swsc/2024002/olm.

Table S1: Explanatory variables trialed within the statistical models. The Swarm data products are available at ftp://swarm-diss.eo.esa.int, the OMNI resolution data are available at https://spdf.gsfc.nasa.gov/pub/data/omni/, the Service International des Indices Géomagnétiques data are available at https://isgi.unistra.fr/geomagnetic_indices.php and the Laboratory for Atmospheric and Space Physics data are available at https://lasp.colorado.edu.

Table S2: The distribution chosen, the transformation applied to the dependent variable, the explanatory variables (EV), parameter estimates (PE) and the uncertainties in the parameter estimates (ΔPE) for the version 1 of the models fitted. An explanation of the terms in the model is given in Table S1.

Table S3: As Table S2, but with separate models in the equatorial region for different times of day.

Table S4: As Table S2, but for version 2 of the models.

References

Aarons J. 1982. Global Morphology of Ionospheric Scintillations. *Proc IEEE* **70(4)**: 360–378. https://doi.org/10.1109/PROC.1982. 12314.

Akasofu S-I. 1996. Search for the "unknown" quantity in the solar wind: A personal account. *J. Geophys. Res.* **101**: 10531–10540. https://doi.org/10.1029/96JA00182.

Appleton E. 1946. Two anomalies in the ionosphere. *Nature* **157(3995)**: 691. https://doi.org/10.1038/157691a0.

Balan N, Liu L, Le H. 2018. A brief review of equatorial ionization anomaly and ionospheric irregularities. *Earth Planet Phys* **2**: 257–275. https://doi.org/10.26464/epp2018025.

Barlow RJ. 1989. *Statistics: A guide to the use of statistical methods in the physical sciences*. Wiley, Chichester, UK. ISBN: 978-0-471-92295-7.

Basu S, Basu S. 1981. Equatorial scintillations – A review. *J Atmos Sol-Terr Phys* **43(5–6)**: 473–489. https://doi.org/10.1016/0021-9169(81)90110-0.

Bezděk A, Sebera J, Klokočník J. 2018. Calibration of Swarm accelerometer data by GPS positioning and linear temperature correction. *Adv Space Res* **62(2)**: 317–325. https://doi.org/10.1016/j.asr.2018.04.041.

Birkeland K. 1913. *The Norwegian aurora polaris expedition 1902-03*, Vols. **I and II**, Aschehoug, Christiania, Norway.

Boyde B, Wood A, Dorrian G, Fallows R, Themens D, Mielich J, et al. 2022. Lensing from small-scale travelling ionospheric disturbances observed using LOFAR. *J Space Weather Space Clim* **12**: 34. https://doi.org/10.1051/swsc/2022030.

Brekke A. 1997. *Physics of the upper polar atmosphere, Wiley-Praxis series in atmospheric physics*. Wiley, Chichester, UK. ISBN: 0471960187.

Buchau J, Reinisch BW, Weber EJ, Moore JG. 1983. Structure and dynamics of the winter polar cap F region. *Radio Sci* **18**: 995–1010. https://doi.org/10.1029/RS018i006p00995.

Cherniak I, Zakharenkova I, Sokolovsky S. 2019. Multi-instrumental observation of storm-induced ionospheric plasma bubbles at equatorial and middle latitudes. *J Geophys Res Space Phys* **124**: 1491–1508. https://doi.org/10.1029/2018JA026309.

Cherniak I, Zakharenkova I. 2016. First observations of super plasma bubbles in Europe. *Geophys Res Lett* **43**: 11137–11145. https://doi.org/10.1002/2016GL071421.

Crowley G. 1996. Critical Review of patches and blobs. In: *Polar Cap Boundary Phenomena, in: URSI Review of Radio Science 1993–1996*, Stone WR (Ed.), published for the International Union of Radio Science, Oxford University Press, pp. 619–648.

De Franceschi G, Alfonsi L, Romano V, Aquino M, Dodson A, Mitchell CN, Spencer P, Wernik AW. 2008. Dynamics of high-latitude patches and associated small-scale irregularities during the October and November 2003 storms. *J Atmos Sol Terr Phys* **70(6)**: 879–888. https://doi.org/10.1016/j.jastp.2007.05.018.

De Franceschi G, Spogli L, Alfonsi L, Romano V, Cesaroni C, Hunstad I. 2019. The ionospheric irregularities climatology over Svalbard from solar cycle 23. *Sci Rep* **9(1)**: 1–14. https://doi.org/10.1038/s41598-019-44829-5.

Dorrian GD, Wood AG, Ronksley A, Aruliah A, Shahtahmassebi G. 2019. Statistical modelling of the coupled F-region ionosphere-thermosphere at high latitude during polar darkness. *J Geophys Res* **124**: 1389–1409. https://doi.org/10.1029/2018JA026171.

Elmas Z, Forte B, Aquino A. 2011. The impact of ionospheric scintillation on the GNSS receiver signal tracking performance and measurement accuracy. In: URSI General Assembly and Scientific Symposium, Istanbul, Turkey, 13–20 August 2011. https://doi.org/10.1109/URSIGASS.2011.6123719.

Fallows RA, Forte B, Astin I, Allbrook T, Arnold A, Wood A, et al. 2020. A LOFAR observation of ionospheric scintillation from simultaneous medium- and large-scale travelling ionospheric disturbances. *J Space Weather Space Clim* **10**: 10. https://doi.org/10.1051/swsc/2020010.

Foster JC. 1984. Ionospheric signatures of magnetospheric convection. *J Geophys Res* **89**: 855–865. https://doi.org/10.1029/JA089iA02p00855.

Francis SH. 1975. Global propagation of atmospheric gravity waves: a review. *J Atmos Terr Phys* **37**: 1011–1030. https://doi.org/10.1016/0021-9169(75)90012-4.

Friis-Christensen E, Lühr H, Hulot G. 2006. Swarm: a constellation to study the Earth's magnetic field. *Earth Planets Space* **58(4)**: 351–358. https://doi.org/10.1186/BF03351933.

Ghobadi H, Spogli L, Alfonsi L, Cesaroni C, Cicone A, Linty N, Romano V, Cafaro M.. 2020. Disentangling ionospheric refraction and diffraction effects in GNSS raw phase through fast iterative filtering technique. *GPS Solut* **24(3)**: 1–13. https://doi.org/10.1007/s10291-020-01001-1.

Hargreaves JK. 1992. *The solar-terrestrial environment, Cambridge atmospheric and space science series*, *Cambridge University Press*, Cambridge. https://doi.org/10.1017/CBO9780511628924.

Hey JS, Parsons SJ, Phillips JW. 1946. Fluctuations in cosmic radiation at radiofrequencies. *Nature* **158**: 247. https://doi.org/10.1038/158234a0.

Hill GE. 1963. Sudden enhancements of F-layer ionization in polar regions. *J Atmos Sci* **20**: 492–497. https://doi.org/10.1175/1520-0469(1963)020<0492:SEOLII>2.0.CO;2.

Hinteregger HE. 1977. EUV flux variation during end of solar cycle 20 and beginning cycle 21, observed from AE-C satellite. *Geophys Res Letts* **4(6)**: 231–234. https://doi.org/10.1029/GL004i006p00231.

Hunsucker RD. 1982. Atmospheric gravity waves generated in the high-latitude ionosphere: A review. *Rev Geophys* **20(2)**: 293–315. https://doi.org/10.1029/RG020i002p00293.

Kil H, Heelis RA. 1998. Global distribution of density irregularities in the equatorial ionosphere. *J Geophys Res Space Phys* **103(A1)**: 407–417. https://doi.org/10.1029/97JA02698.

Kinrade J, Mitchell CN, Smith ND, Ebihara Y, Weatherwax AT, Bust GS. 2013. GPS phase scintillation associated with optical auroral emissions: first statistical results from the geographic South Pole. *J Geophys Res* **118**: 2490–2502. https://doi.org/10.1002/jgra.50214.

Koskinen HE, Tanskanen EI. 2002. Magnetospheric energy budget and the epsilon parameter. *J Geophys Res Space Phys* **107(A11)**: SMP 42-1–SMP 42-10. https://doi.org/10.1029/2002JA009283.

Kotova D, Jin Y, Spogli L, Wood AG, Urbar J, Rawlings JT, Whittaker IC, Alfonsi L, Clausen LBN, Høeg P, Miloch WJ. 2023. Electron density fluctuations from Swarm as a proxy for ground-based scintillation data: a statistical perspective. *Adv Space Res* **72**: 5399–5415. https://doi.org/10.1016/j.asr.2022.11.042.

Kotova D, Jin Y, Miloch W. 2022. Interhemispheric variability of the electron density and derived parameters by the Swarm satellites during different solar activity. *J Space Weather Space Clim* **12**: 12. https://doi.org/10.1051/swsc/2022007.

Jenner L, Wood AG, Dorrian GD, Oksavik K, Yeoman TK, Fogg A. 2020. Plasma density gradients at the edge of polar ionospheric holes: the absence of phase scintillation. *Ann. Geophys.* **38**: 575–590. https://doi.org/10.5194/angeo-38-575-2020.

Jin Y, Kotova D, Xiong C, Brask SM, Clausen LBN, Kervalishvili G, Stolle C, Miloch WJ. 2022. Ionospheric plasma irregularities – IPIR – Data product based on data from the swarm satellites. *J Geophys Res Space Phys* **127(4)**: e2021JA030183. https://doi.org/10.1029/2021JA030183.

Jin Y, Xiong C, Clausen L, Spicher A, Kotova D, Brask S, Kervalishili G, Stolle C, Miloch WJ. 2020. Ionospheric plasma irregularities based on in situ measurements from the Swarm satellites. *J Geophys Res Space Phys* **125(7)**: e2020JA028103. https://doi.org/10.1029/2020JA028103.

Jin Y, Spicher A, Xiong C, Clausen LBN, Kervalishvili G, Stolle C, Miloch WJ. 2019. Ionospheric plasma rregularities characterized by the Swarm satellites: statistics at high latitudes. *J Geophys Res Space Phys* **124**: 1262–1282. https://doi.org/10.1029/2018JA026063.

Jin Y, Moen J, Miloch WJ. 2014. GPS scintillation effects associated with polar cap patches and substorm auroral activity: direct comparison. *J Space Weather Space Clim* **4**: A23. https://doi.org/10.1051/SWSC/2014019.

Jin Y, Moen JI, Miloch WJ, Clausen LBN, Oksavik K. 2016. Statistical study of the GNSS phase scintillation associated with two types of auroral blobs. *J Geophys Res Space Phys* **121**: 4679–4697. https://doi.org/10.1002/2016JA022613.

Jin Y, Oksavik K. 2018. GPS scintillations and losses of signal lock at high latitudes during the 2015 St. Patrick's Day storm. *J Geophys Res Space Phys* **123**: 7943–7957. https://doi.org/10.1029/2018JA025933.

Landerer FW, Flechtner FM, Save H, Webb FH, Bandikova T, Bertiger WI, et al. 2020. Extending the global mass change data record: GRACE Follow-On instrument and science data performance. *Geophys Res Lett* **47**: e2020GL088306. https://doi.org/10.1029/2020GL088306.

Li G, Ning B, Otsuka Y, Abdu MA, Abadi P, Liu Z, et al. 2021. Challenges to equatorial plasma bubble and ionospheric scintillation short-term forecasting and future aspects in east and southeast Asia. *Surv Geophys* **42(1)**: 201–238. https://doi.org/10.1007/s10712-020-09613-5.

Liemohn MW, Shane AD, Azari AR, Petersen AK, Swiger BM, Mukhopadhyay A. 2021. RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *J Atmos Sol Terr Phys* **218**: 105624. https://doi.org/10.1016/j.jastp.2021.105624.

Lockwood M, Carlson HC Jr. 1992. Production of polar cap electron density patches by transient magnetopause reconnection. *Geophys Res Lett* **19**: 1731–1734. https://doi.org/10.1029/92GL01993.

McCaffrey AM, Jayachandran PT. 2019. Determination of the refractive contribution to GPS phase "scintillation". *J Geophys Res Space Phys* **124(2)**: 1454–1469. https://doi.org/10.1029/2018JA025759.

McClure JP, Hanson WB, Hoffman JH. 1977. Plasma bubbles and irregularities in the equatorial ionosphere. *J Geophys Res* **82(19)**: 2650–2656. https://doi.org/10.1029/JA082i019p02650.

McCullagh P, Nelder JA. 1983. *Generalized linear models. CRC monographs on statistics and applied probability*, Chapman and Hall, London. ISBN 10:0412238500.

Mitchell CN, Alfonsi L, De Franceschi G, Lester M, Romano V, Wernik AW. 2005. GPS TEC and scintillation measurements from the polar ionosphere during the October 2003 storm. *Geophys Res Lett* **32**: L12S03. https://doi.org/10.1029/2004GL021644.

Mott-Smith H, Langmuir I. 1926. The theory of collectors in gaseous discharges. *Phys Rev* **28**: 27. https://doi.org/10.1103/PhysRev.28.727.

Newell PT, Sotirelis T, Liou K, Meng C-I, Rich FJ. 2007. A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables. *J Geophys Res* **112**: A01206. https://doi.org/10.1029/2006JA012015.

Pedersen PO. 1927. *Propagation of radio waves*. Danmarks Natur. Samf, Copenhagen.

Pedersen TR, Fejer BG, Doe RA, Weber EJ. 2000. An incoherent scatter radar technique for determining two-dimensional horizontal ionization structure in polar cap F region patches. *J Geophys Res* **105**: 10637–10655. https://doi.org/10.1029/1999JA000073.

Prikryl P, Jayachandran PT, Chadwick R, Kelly TD. 2015. Climatology of GPS phase scintillation at northern high latitudes for the period from 2008 to 2013. *Ann Geophys* **33**: 531–545. https://doi.org/10.5194/angeo-33-531-2015.

Pryse SE, Wood AG, Middleton HR, McCrea IW, Lester M. 2006. Reconfiguration of polar cap plasma in the magnetic midnight sector. *Ann Geophys* **24**: 2201–2208. https://doi.org/10.5194/angeo-24-2201-2006.

Rajesh PK, Lin CCH, Lin JT, Lin CY, Liu JY, Matsuo T, et al. 2022. Extreme poleward expanding super plasma bubbles over Asia-Pacific region triggered by Tonga volcano eruption during the recovery-phase of geomagnetic storm. *Geophys Res Lett* **49**: e2022GL099798. https://doi.org/10.1029/2022GL099798.

Rees MH. 1989. *Physics and chemistry of the upper atmosphere, Cambridge atmospheric and space science series*. Cambridge University Press, Cambridge. ISBN: 9780521368483.

Rishbeth H, Setty CSGK. 1961. The F-layer at sunrise. *J Atmos Terr Phys* **21**: 263–276. https://doi.org/10.1016/0021-9169(61)90205-7.

Rishbeth H, Mendillo M. 2001. Patterns of F2-layer variability. *J Atmos Sol Terr Phys* **63**: 1661–1680. https://doi.org/10.1016/S1364-6826(01)00036-0.

Rishbeth H. 1971. Polarization fields produced by winds in the equatorial F-region. *Planet Space Sci* **19(3)**: 357–369. https://doi.org/10.1016/0032-0633(71)90098-5.

Rodger AS, Pinnock M, Dudeney JR, Baker KB, Greenwald RA. 1994. A new mechanism for polar patch formation. *J Geophys Res* **99(A4)**: 6425–6436. https://doi.org/10.1029/93JA01501.

Schwemer G. 2000. General linear models for multicenter clinical trials. *Control Clin Trials* **21(1)**: 21–29. https://doi.org/10.1016/s0197-2456(99)00035-5.

Smith AM, Mitchell CN, Watson RJ, Meggs RW, Kintner PM, Kauristie K, Honary F. 2008. GPS scintillation in the high arctic associated with an auroral arc. *Space Weather* **6**: S03D01. https://doi.org/10.1029/2007SW000349.

Sojka JJ, Bowline MD, Schunk RW, Decker DT, Balladares CE, Sheehan R, Anderson DN, Heelis RA. 1993. Modelling polar-cap F-region patches using time-varying convection. *Geophys Res Lett* **20**: 1783–1786. https://doi.org/10.1029/93GL01347.

Spogli L, Jin Y, Urbář J, Wood AG, Donegan-Lawley EE, Clausen LBN, et al. 2024. Statistical models of the variability of plasma in the topside ionosphere: 2: Performance assessment. *J Space Weather Space Clim*. https://doi.org/10.1051/swsc/2024003.

Spogli L, Ghobadi H, Cicone A, Alfonsi L, Cesaroni C, Linty N, Romano V, Cafaro M. 2021. Adaptive phase detrending for GNSS scintillation detection: A case study over Antarctica. *IEEE Geosci Remote Sens Lett* **19**: 1–5. https://doi.org/10.1109/LGRS.2021.3067727.

Spogli L, Alfonsi L, De Franceschi G, Romano V, Aquino MHO, Dodson A. 2009. Climatology of GPS ionospheric scintillations over high and mid-latitude European regions. *Ann Geophys* **27**: 3429–3437. https://doi.org/10.5194/angeo-27-3429-2009.

Sun W, Kuriakose AK, Li G, Li Y, Zhao X, Hu L, et al. 2022. Unseasonal super ionospheric plasma bubble and scintillations seeded by the 2022 Tonga Volcano Eruption related perturbations. *J Space Weather Space Clim* **12**: 25. https://doi.org/10.1051/swsc/2022024.

Tapping KF. 2013. The 10.7 cm solar radio flux (F10.7). *Space Weather* **11**: 394–406. https://doi.org/10.1002/swe.20064.

Themens DR, Watson C, Žagar N, Vasylkevych S, Elvidge S, McCaffrey A, et al. 2022. Global propagation of ionospheric disturbances associated with the 2022 Tonga volcanic eruption. *Geophys Res Lett* **49**: e2022GL098158. https://doi.org/10.1029/2022GL098158.

Urbar J, Spogli L, Cicone A, Clausen LBN, Jin Y, Wood AG, et al. 2022. Multi-scale response of the high-latitude topside ionosphere to geospace forcing. *Adv Space Res* **72**: 5490–5502. https://doi.org/1010.1016/j.asr.2022.06.045.

Valladares CE, Decker DT, Sheehan R, Anderson DN, Bullett T, Reinisch BW. 1998. Formation of polar cap patches associated with north-to-south transitions of the interplanetary magnetic field. *J Geophys Res* **103(A7)**: 14657–14670. https://doi.org/10.1029/97JA03682.

Valladares CE, Basu S, Buchau J, Friis-Christensen E. 1994. Experimental evidence for the formation and entry of patches into the polar cap. *Radio Sci* **29(1)**: 167–194. https://doi.org/10.1029/93RS01579.

van den IJssel J, Doornbos E, Iorfida E, March G, Siemes C, Montenbruck O. 2020. Thermosphere densities derived from Swarm GPS observations. *Adv Space Res* **65(7)**: 1758–1771. https://doi.org/10.1016/j.asr.2020.01.004.

Walker IK, Moen J, Kersley L, Lorentzen DA. 1999. On the possible role of cusp/cleft precipitation in the formation of polar-cap patches. *Ann Geophys* **17**: 1298–1305. https://doi.org/10.1007/s00585-999-1298-4.

Weber EJ, Klobuchar JA, Buchau J, Carlson HC Jr, Livingston RC, de la Beaujadiere O, McCready M, Moore JG, Bishop GJ. 1986. Polar cap F layer patches: Structure and dynamics. *J Geophys Res* **91**: 12121–12129. https://doi.org/10.1029/JA091iA11p12121.

Weber EJ, Buchau J, Moore JG, Sharber JR, Livingston RC, Winningham JD, Reinisch BW. 1984. F layer ionisation patches in the polar cap. *J Geophys Res* **89**: 1683–1694. https://doi.org/10.1007/s00585-999-1298-4.

Wernik AW, Secan JA, Fremouw EJ. 2003. Ionospheric irregularities and scintillation. *Adv Space Res* **31**: 971–981. https://doi.org/10.1016/S0273-1177(02)00795-0.

Wood AG, Alfonsi L, Clausen LBN, Jin Y, Spogli L, Urbar J, et al. 2022. Variability of ionospheric plasma: results from the ESA Swarm Mission. *Space Sci Rev* **218**: 52. https://doi.org/10.1007/s11214-022-00916-0.

Wood AG, Mountain L, Connors R, Maher M, Ropkins K. 2013. Updating outdated predictive accident models. *Accid Anal Prev* **55**: 48–53. https://doi.org/10.1016/j.aap.2013.02.028.

Wood AG, Pryse SE. 2010. Seasonal influence on polar cap patches in the high-latitude nightside ionosphere. *J Geophys Res* **115**: A07311. https://doi.org/10.1029/2009JA014985.

Woodman RF, La Hoz C. 1976. Radar observations of F region equatorial irregularities. *J Geophys Res Space Phys* **81(31)**: 5447–5466. https://doi.org/10.1029/JA081i031p05447.

Wright JW. 1963. The F region seasonal anomaly. *J Geophys Res* **68**: 4379–4381. https://doi.org/10.1029/JZ068i014p04379.

Wright CJ, Hindley NP, Alexander MJ, Barlow M, Hoffmann L, Mitchell CN, et al. 2022. Surface-to-space atmospheric waves from Hunga Tonga-Hunga Ha'apai eruption. *Nature* **609**: 741–746. https://doi.org/10.1038/s41586-022-05012-5.

Zhang Q-H, Ma Y-Z, Jayachandran PT, Moen J, Lockwood M, Zhang Y-L, et al. 2017. Polar cap hot patches: Enhanced density structures different from the classical patches in the ionosphere. *Geophys Res Lett* **44**: 8159–8167. https://doi.org/10.1002/2017GL073439.

Cite this article as: Wood AG, Donegan-Lawley EE, Clausen LBN, Spogli L, Urbář J, et al. 2024. Statistical models of the variability of plasma in the topside ionosphere: 1. Development and optimisation. *J. Space Weather Space Clim*. **14**, 7. https://doi.org/10.1051/swsc/2024002.