Space Weather

RESEARCH ARTICLE
10.1029/2018SW002018

New Capabilities for Prediction of High-Latitude Ionospheric Scintillation: A Novel Approach With Machine Learning

Ryan M. McGranaghan[1,2], Anthony J. Mannucci[2], Brian Wilson[2], Chris A Mattmann[2], and Richard Chadwick[3]

[1]University Corporation for Atmospheric Research, Boulder, CO, USA, [2]NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA, [3]Department of Physics, University of New Brunswick, Fredericton, New Brunswick, Canada

Key Points:
- Support vector machine yields predictive performance roughly double that of persistence; the benefit increases with longer prediction times
- Electron precipitation information is critical to the prediction of high-latitude phase scintillation
- A benchmark is established for high-latitude ionospheric phase scintillation using the robust true skill score (TSS) metric

Correspondence to:
R. M. McGranaghan,
ryan.mcgranaghan@jpl.nasa.gov

Citation:
McGranaghan, R. M., Mannucci, A. J., Wilson, B. D., Mattmann, C. A., & Chadwick, R. (2018). New capabilities for prediction of high-latitude ionospheric scintillation: A novel approach with machine learning. *Space Weather*, *16*, 1817–1846. https://doi.org/10.1029/2018SW002018

Received 17 JUL 2018
Accepted 5 OCT 2018
Accepted article online 12 OCT 2018
Published online 18 NOV 2018

Abstract As societal dependence on transionospheric radio signals grows, space weather impact on these signals becomes increasingly important yet our understanding of the effects remains inadequate. This challenge is particularly acute at high latitudes where the effects of space weather are most direct and no reliable predictive capability exists. We take advantage of a large volume of data from Global Navigation Satellite Systems (GNSS) signals, increasingly sophisticated tools for data-driven discovery, and a machine learning algorithm known as the support vector machine (SVM) to develop a novel predictive model for high-latitude ionospheric phase scintillation. This work, to our knowledge, represents the first time an SVM model has been created to predict high-latitude phase scintillation. We use the true skill score to evaluate the SVM model and to establish a benchmark for high-latitude ionospheric phase scintillation prediction. The SVM model significantly outperforms persistence (i.e., current and future scintillation are identical), doubling the predictive skill according to the true skill score for a 1-hr lead time. For a 3-hr lead time, persistence is comparable to a random chance prediction, suggesting that the *memory* of the ionosphere in terms of high-latitude plasma irregularities is on the order of, or shorter than, a few hours. The SVM model predictive skill only slightly decreases between the 1- and 3-hr predictive tasks, pointing to the potential of this method. Our findings can serve as a foundation on which to evaluate future predictive models, a critical development toward the resolution of space weather impact on transionospheric radio signals.

Plain Language Summary Society is increasingly dependent on radio signals, particularly those from the Global Navigation Satellite Systems (GNSS), and the technologies (e.g., navigation and financial transactions) that they enable. The integrity and reliability of these signals is threatened by their travel from the GNSS satellites to the ground, which includes passage through a charged region between 100 and 1,000 km known as the ionosphere. Disturbances to the ionosphere from solar energy, or space weather, cause variations in GNSS signals that adversely affect the dependent systems and technologies. Currently, the effect of the ionosphere on these signals cannot be reliably predicted, and the challenge is particularly important at latitudes above 45° where space weather impacts are most direct. We have compiled a large volume of data from the regions important to space weather (i.e., from the Sun to the Earth) to develop a novel machine learning model capable of skillfully predicting disruptions to GNSS signals at high latitudes. To our knowledge, this model is the first of its kind. We find that the new model is capable of more accurate predictions than current methods and position this model as a benchmark on which future predictive models can be measured.

1. Introduction

Irregularities in the density of the charged region of the upper atmosphere between ~100- and 1,000-km altitude—the ionosphere—cause rapid radio signal phase and intensity fluctuations in the frequency range between 100 MHz and 4 GHz (Aarons & Basu, 1994; Basu et al., 1988; Kintner, 2001). These fluctuations are referred to as scintillation, and their study, particularly at high latitudes and in the context of recent advances in machine learning methodologies, is in its infancy.

Ionospheric scintillation is increasingly important as (1) our society becomes more dependent on Global Navigation Satellite Systems (GNSS) signals, which are critically affected by the ionosphere (Kintner et al., 2007), and (2) proliferation of technology and access to space drives a greater reliance on transionospheric signals

(Tzelepis & Carreno, 2016). Despite the frequent and deleterious effects of the ionosphere on radio signals there is a lack of awareness of, and appreciation for, mitigating these impacts on a given GNSS-dependent service. Due to the absence of robust prediction models, users are often unaware that disruptions should be attributed to ionospheric disturbances. This fact, exacerbated by the increasing demand on GNSS applications (e.g., Sadlier et al., 2017), motivates the pressing need for new predictive capabilities.

Though powerful new understanding has been created based on improved models both of the ionosphere (Wernik et al., 2003) and for signal propagation through irregularities (Chartier et al., 2016; Deshpande et al., 2016; Forte, 2012; Norman et al., 2016) most new knowledge is *incremental*, coming in two forms: (1) highly specific cases without clear broad connection/implications or (2) highly generalized circumstances without application to specific individual situations. For example, Chartier et al. (2016) provided an extensive analysis of an event from 20:03 to 20:07 UT on 17 October 2013 during which several observing systems provided constraints from which a scintillation model could be tuned. Their study contributed new understanding of phase scintillation; however, how well the results generalize is unknown. Despite the importance of these advances, they do not constitute a comprehensive understanding capable of unifying the general and the specific. Both approaches are limited in predictive capability. The former only applies to very specific circumstances, while the latter applies only in a qualitative sense and cannot accurately inform the individual cases that are most important (i.e., disturbed ionospheric conditions). The result is a lack of effective ionospheric scintillation predictive capability, and the need is particularly acute at high latitudes.

While prediction of scintillation at low latitudes is, in many respects, a more tractable problem given the predominance of repeatable (i.e., predictable) physics associated with the equatorial ionization anomaly (Anderson, 1973; Hanson & Moffett, 1966; Muella et al., 2010) and has received more attention as a result (Redmon et al., 2010; Uwamahoro & Habarulema, 2015), prediction of scintillation at high latitudes involves additional complexity due to more direct space weather connections (Cowley, 2013).

The body of work to predict high-latitude ionospheric scintillation is very limited. The simplest approach is persistence prediction or predicting that the scintillation conditions in the future will be identical to the conditions currently. The accuracy of a persistence prediction will decrease with increasing prediction time. To be considered skillful in any way, any new prediction model must outperform persistence. Another approach is climatological prediction, in which scintillation statistics are accumulated over many years, perhaps binned according to solar wind or geomagnetic activity variables, and future scintillation occurrence is predicted based on the statistical occurrence for a given date, time, and location. An upgrade to the high-latitude portion of the WideBand MODel of ionospheric scintillation (Secan, 1995), known as SCINTMOD (Secan et al., 1997), is capable of providing climatological predictions of high-latitude scintillation. SCINTMOD produces predictions of ionospheric scintillation based on *F*-region electron density climatology with variations due to sunspot number, the *Kp* index, latitude, local time, longitude, and season. A variant on climatological prediction utilizing probabilistic relationships derived for repeatable solar conditions such as coronal mass ejections and corotating interaction regions was introduced for geospace applications by McPherron and Siscoe (2004). This approach is derived from the air mass climatology paradigm in meteorological prediction. Prikryl et al. (2012, 2013) adapted this approach for high-latitude scintillation using Canadian High Arctic Ionospheric Network (CHAIN) data, though this applies only to scintillation prediction during coronal mass ejections and corotating interaction regions and has only been experimentally applied to a select few case studies (i.e., not available for widespread application).

The data-driven (e.g., machine learning) approach to prediction, on the other hand, relies on a large collection of input data with an associated label describing the variable to be predicted and attempts to automatically extract the relationships between these data. Despite limited application to low-latitude scintillation (Habarulema et al., 2011; Lima et al., 2015; Rezende et al., 2009; Uwamahoro & Habarulema, 2015), machine learning prediction techniques are largely unexplored for middle- and high-latitude (defined here to mean $\geq$45° magnetic latitude [MLAT]) ionospheric scintillation. To our knowledge we provide the first investigation of machine learning prediction for middle- and high-latitude ionospheric phase scintillation.

At latitudes poleward of 45° characteristics of phase variations/scintillation are not well understood. Though climatological studies of high-latitude scintillation have improved the understanding of these phenomena (Alfonsi et al., 2011; Jiao et al., 2013; Spogli et al., 2009), their use for prediction is limited. In fact, *high-latitude scintillation prediction lacks a reliable benchmark* from which to evaluate future improvements.

While the proliferation of transionospheric radio signals and technologies dependent on them have produced a dire need to understand and predict scintillation, it has also created a much wider data set through which to study, understand, and, ultimately, predict the phenomenon. The operation of the United States' Global Positioning System (http://www.igs.org/) constellation since 1993 coupled with the advent of Russian (Globalnaya Navigazionnaya Sputnikovaya Sistema), Chinese (Beidou, http://www.beidou.gov.cn), and European (Galileo, http://www.gsa.europa.eu/galileo/programme) systems and the proliferation of ground-based receivers and networks of receivers (e.g., the International GNSS Service high-latitude network [http://www.igs.org/, Cherniak et al., 2014], CHAIN [http://chain.physics.unb.ca/chain/], Greenland Global Positioning System Network [http://www.polar.dtu.dk/english/Research/Facilities/GNET], Istituto Nazionale di Geofisica e Vulcanologia Electronic Space Weather Upper Atmosphere [http://www.eswua.ingv.it/ingv/home.php?res=1024], GNSS Earth Observation NETwork [http://datahouse1.gsi.go.jp/terras/terras_english.html]) provide a vast and powerful new data set through which ionospheric scintillation and, more generally, space weather can be studied (Beutler et al., 1999; Ren et al., 2016; Rizos et al., 2013). These data provide information at higher cadence and over a larger portion of the globe than any other single data set and are the premier remote sensing tools to facilitate new understanding of space weather phenomena (Coster & Komjathy, 2008). GNSS data are voluminous (on the order of terabytes when considering data from the mid-1990s to now and taking into account the growing number of ground-based receivers) and heterogeneous (different satellite systems, receiver technologies, and archived in a variety of locations and formats). Appropriate utilization of these data can potentially revolutionize the study of space weather.

The presence of large volumes of underutilized data motivates the investigation of machine learning approaches. Machine learning here is broadly defined as any approach that allows a computer system to learn from experience introduced in the form of data samples. Our definition of machine learning encapsulates a broad range of approaches, including linear regression, clustering, information theory, statistical modeling, and neural networks, to name a few. There are, in general, three keys to successful machine learning: (1) availability of a large volume of high-quality data, (2) a well-defined task, and (3) adequate computational resources. The advent of GNSS signals coupled with long histories of ground-based GNSS signal observation and a robust data quality control process (detailed in section 2) carried out for this work address the first requirement. To address the second requirement, we study the explicit task of predicting the occurrence of ionospheric scintillation at any given location with 1-hr lead time given input solar wind and geospace data (further detailed in section 3). Finally, to satisfy the computational demand, this work takes advantage of increased availability of computational resources (e.g., high-performance computing) that have become commonplace in the digital age (Duderstadt, 2001).

The following three questions (and the sections of this paper where they are addressed) motivate this work:

- What data are most important to determine future high-latitude phase scintillation activity? (section 3.2)
- Can we establish a benchmark for prediction of high-latitude ionospheric scintillation *at spatial resolutions commensurate with user needs*? (section 5)
- To what extent can current understanding of space weather phenomena guide the improvement of machine learning models and what future research paths does this outline? (section 6.1)

The emphasis and importance of this paper is to use an *explainable* machine learning technique known as support vector machines (SVMs) to establish new relationships between observed solar wind, geomagnetic activity, and ionospheric behavior and future phase scintillation occurrence in the high-latitude ionosphere without attempting to explain the complex and numerous physical mechanisms giving rise to ionospheric irregularities and the physical relationships that cause these irregularities to lead to scintillation for a given circumstance (i.e., at a given location with specific contextual conditions). We refer to the *explainability* of SVMs as the quality by which links are explicitly identified between the scintillation prediction and the input variables (for more information on *explainable* methods refer to, for example, the Defense Advance Research Projects Agency project, Explainable Artificial Intelligence, https://www.darpa.mil/program/explainable-artificial-intelligence). We choose SVM due to its high accuracy, capability to handle high-dimensional data, and flexibility to model diverse data (further justification is provided in section 3.3). The results of this work provide a benchmark for future prediction models through the use of the true skill score (TSS; Bloomfield et al., 2012) as the model evaluation metric.

Therefore, the contributions of this paper are threefold:

Table 1
Input Features and Details[a,b,c]

Feature	*F* score	Description
DOY	1.10	Day of year
UT	1.06	Universal time (seconds into the day)
azimuth (°)	1.31	Azimuth of receiver-satellite line of sight
elevation (°)	1.07	Elevation of receiver-satellite line of sight
B_z—30 min (nT)	1.53	Solar wind interplanetary magnetic field (IMF) *z* component 30 min prior to observation
B_z—15 min (nT)	1.53	Solar wind IMF *z* component 15 min prior to observation
B_z—0 min (nT)	1.54	Solar wind IMF *z* component at time of observation
B_y—30 min (nT)	1.25	Solar wind IMF *y* component 30 min prior to observation
B_y—15 min (nT)	1.25	Solar wind IMF *y* component 15 min prior to observation
B_y—0 min (nT)	1.24	Solar wind IMF *y* component at time of observation
V_{sw}—30 min (km/s)	1.15	Solar wind velocity 30 min prior to observation
V_{sw}—15 min (km/s)	1.15	Solar wind velocity 15 min prior to observation
V_{sw}—0 min (km/s)	1.16	Solar wind velocity at time of observation
P_{sw}—30 min (nPa)	1.56	Solar wind pressure 30 min prior to observation
P_{sw}—15 min (nPa)	1.55	Solar wind pressure 15 min prior to observation
P_{sw}—0 min (nPa)	1.56	Solar wind pressure at time of observation
AE—30 min (nT)	2.00	Auroral electrojet index 30 min prior to observation
AE—15 min (nT)	2.10	Auroral electrojet index 15 min prior to observation
AE—0 min (nT)	2.17	Auroral electrojet index at time of observation
SymH—30 min (nT)	1.59	Sym-H index 30 min prior to observation
SymH—15 min (nT)	1.60	Sym-H index 15 min prior to observation
SymH—0 min (nT)	1.61	Sym-H index at time of observation
Clock angle—30 min (°)	0.88	Solar wind IMF clock angle 30 min prior to observation
Clock angle—15 min (°)	0.87	Solar wind IMF clock angle 15 min prior to observation
Clock angle—0 min (°)	0.87	Solar wind IMF clock angle at time of observation
Newell CF—30 min ($m/s^{(4/3)}$ $T^{(2/3)}$)	1.99	Newell et al. (2007) coupling function 30 min prior to observation
Newell CF—15 min ($m/s^{(4/3)}$ $T^{(2/3)}$)	2.01	Newell et al. (2007) coupling function 15 min prior to observation
Newell CF—0 min ($m/s^{(4/3)}$ $T^{(2/3)}$)	2.04	Newell et al. (2007) coupling function at time of observation
Borovsky CF—30 min (nT·km/s)	1.11	Borovsky (2013) coupling function 30 min prior to observation
Borovsky CF—15 min (nT·km/s)	1.15	Borovsky (2013) coupling function 15 min prior to observation
Borovsky CF—0 min (nT·km/s)	1.20	Borovsky (2013) coupling function at time of observation
Kp (dimensionless)	1.81	*Kp* index
F107 (sfu = 10^{-22} $W{\cdot}m^{-2}$ Hz^{-1})	1.10	10.7-cm wavelength solar flux
OVATION diffuse eflux ($erg{\cdot}cm^{-2}{\cdot}^{-1}$)	1.92	OVATION Prime auroral precipitation model diffuse electron energy flux
OVATION mono eflux ($erg{\cdot}cm^{-2}{\cdot}s^{-1}$)	2.23	OVATION Prime auroral precipitation model monoenergetic electron energy flux
OVATION wave eflux ($erg{\cdot}cm^{-2}{\cdot}s^{-1}$)	2.47	OVATION Prime auroral precipitation model wave electron energy flux

Table 1 (*continued*)

OVATION diffuse nflux (#·cm^{-2}·s^{-1})	1.92	OVATION Prime auroral precipitation model diffuse electron number flux
OVATION mono nflux (#·cm^{-2}·s^{-1})	2.25	OVATION Prime auroral precipitation model monoenergetic electron number flux
OVATION wave nflux (#·cm^{-2}·s^{-1})	2.33	OVATION Prime auroral precipitation model wave electron number flux
AACGM latitude (°)	1.42	Altitude-adjusted corrected geomagnetic latitude
AACGM longitude (°)	1.37	Altitude-adjusted corrected geomagnetic longitude
cos(AACGM local time) (rad)	1.02	Cosine of the altitude-adjusted corrected geomagnetic local time
sin(AACGM local time) (rad)	1.01	Sine of the altitude-adjusted corrected geomagnetic local time
geographic latitude (°)	1.47	Geographic latitude
geographic longitude (°)	1.37	Geographic longitude
TEC at current time (TECU)	1.11	Total electron content at time of observation
dTEC 0 min to 15 s, to 0 min to 0 s (TECU)	1.94	Rate of change of the total electron content over the 15 s prior to observation
SI (dimensionless)	1.23	GNSS signal spectral index
Spectral slope (dimensionless)	3.04	GNSS signal spectral slope
S_4 projected to vertical (dimensionless)	1.30	Amplitude scintillation index projected to vertical
σ_Φ projected to vertical (rad)	3.33	Phase scintillation index projected to vertical

Note. GNSS = Global Navigation Satellite Systems; OVATION = Oval Variation, Assessment, Tracking, Intensity, and Online Nowcasting; SI = scintillation index.
[a]Red text coloring refers to solar wind data. [b]Blue text coloring refers to geomagnetic activity and OVATION Prime particle precipitation data. [c]Green text coloring refers to ionospheric data obtained from Canadian High Arctic Ionospheric Network data.

- Curate a new database for use with data-driven techniques to better understand and predict high-latitude phase scintillation and that can be flexibly used to fulfill various space weather user needs.
- Provide new high-latitude phase scintillation prediction capabilities based on machine learning approaches.
- Create a benchmark for high-latitude phase scintillation prediction.

The rest of the paper is outlined as follows: Sections 2 and 3, respectively, first explicitly address the first two key components of a successful machine learning approach: data input and predictive task definition as well as justification for the use of SVM. We then present the results and discussion in sections 5 and 6, respectively. Given the massive exploration space afforded by the GNSS database that we compile, we then discuss the compelling future work in section 6.1. Finally, we provide the conclusions and broader implications in section 7.

2. Data

The objective of this paper is the prediction of high-latitude ionospheric scintillation at specific locations and times, with lead times of a few hours. Our machine learning approach relies on two elements: (1) *features*, or input variables, that contain enough information to explain the diversity of scintillation behavior, and can, therefore, collectively act as a suitable predictor and (2) the corresponding scintillation data at a later time equal to the prediction lead time to be used for training and validation. We take advantage of openly available solar wind, geomagnetic activity, particle precipitation, and ionospheric GNSS data to develop the predictive model and compose the machine learning database. These data are chosen in an attempt to incorporate information from across the solar-terrestrial system (from the Sun, through interplanetary space, into the magnetosphere, and extending down to the upper atmosphere).

2.1. Solar Wind and Geomagnetic Activity Data

We use 5-min resolution solar wind and geomagnetic activity data from NASA's Coordinated Data Analysis Web (CDAWeb, https://cdaweb.sci.gsfc.nasa.gov/). These data contain measurements from multiple space-

craft, accounting for estimated spacecraft-to-magnetopause propagation times. The text colors red and blue in Table 1 respectively detail the solar wind and geomagnetic activity variables used.

2.2. Particle Precipitation Data

Due to the importance of particle precipitation to ionospheric scintillation (Mrak et al., 2017; Semeter et al., 2016, 2017; Zou et al., 2015), we also incorporate data from the Oval Variation, Assessment, Tracking, Intensity, and Online Nowcasting (OVATION) Prime model (Newell et al., 2010) as potentially important input for prediction. OVATION Prime (freely available at http://sourceforge.net/projects/ovation-prime/) provides statistical distributions of particle precipitation in the middle- and high-latitude ionosphere and was created from 11 years (roughly 50 satellite years) of observations from the Defense Meteorological Satellite Program satellites. The model provides number ($\#\cdot cm^{-2}\cdot s^{-1}$) and energy ($ergs\cdot cm^{-2}\cdot s^{-1}$) flux of diffuse, monoenergetic, and broadband electrons and diffuse ions and is driven by solar wind parameters via the Newell coupling function (Newell et al., 2007). OVATION Prime allows precipitation information to be obtained for any location at any time, which is not currently possible through direct precipitation observations and is considered a useful predictive tool (Machol et al., 2012). It is, therefore, deemed important and appropriate for this work. We use the electron precipitation output in this work and interpolate the global OVATION Prime maps to the location of the appropriate observation for use in the machine learning model. Blue text coloring in Table 1 details the geomagnetic activity data, which includes the OVATION Prime particle precipitation data.

2.3. CHAIN Data

To obtain ionospheric information, including phase scintillation/variation, we utilize GNSS data from the CHAIN (Jayachandran et al., 2009). CHAIN consists of 25 specialized GNSS receivers distributed throughout Canada, covering the auroral region, polar cap, and ionospheric cusp (http://chain.physics.unb.ca/chain/). Two types of receivers are included in the network: NovAtel GSV4004B (Dierendonck & Arbesser-Rastburg, 2001) and Septentrio PolaRxS (http://chain.physics.unb.ca/chain/pages/gps/\#PolaRxS). Table 2 provides details of the CHAIN stations. The data we use are subject to the following considerations: (1) We use only PolaRxS receiver data to avoid potential biases with GSV4004B receivers data (16 stations total); (2) data from some stations are not regularly available via the CHAIN file transfer protocol server (R. Chadwick, University of New Brunswick, personal communication, October 2017); (3) Sachs Harbour and Taloyoak station data are unavailable until late 2016; and (4) Kuglugtuk station data are removed due to known bias in the phase data (R. Chadwick, University of New Brunswick, personal communication, October 2017). Each of these caveats are detailed in Table 2. We analyze 2 years of data throughout 2015–2016. Data availability for each CHAIN station are provided by the CHAIN network (http://chain.physics.unb.ca/chain/pages/data_availability).

Details of the CHAIN data processing by which GNSS signals are used to generate ionospheric data products are provided in depth by Jayachandran et al. (2009) and Watson et al. (2016a, 2016b, and references therein) and are not repeated here. We instead focus on the additional data quality control and treatments that we apply. Table 3 provides an overview of these treatments and their rationale. For each CHAIN receiver observation we obtain the following data: location data of the receiver at time of observation (both geographic and altitude-adjusted corrected geomagnetic [AACGM] coordinates [Shepherd, 2014]), azimuth and elevation between the receiver and the GNSS satellite, total electron content (TEC) calculated from the dual L1 and L2 frequency signals, differential TEC calculated as the difference between the TEC value 15 s prior and that at the current observation time, the scintillation index (SI) as defined in section 3.1.9 of Septentrio (2015), the slope of the phase spectral density function (spectral slope, p; e.g., Carrano & Rino, 2016), amplitude SI (S_4), and phase SI (σ_Φ). S_4 is defined as the standard deviation of the 50-Hz raw signal power normalized to the average signal power over the previous minute, while σ_Φ is defined as the standard deviation of the 50-Hz detrended carrier phase averaged over the previous minute. Calculation of S_4 and σ_Φ for the PolaRxS receivers are detailed in Septentrio (2015) and are provided directly in the CHAIN data products. CHAIN receivers collect raw amplitude and phase data at 50 Hz, and the scintillation indices are calculated on 3,000 samples, yielding a 1-min temporal resolution in this study. The ionospheric data, including each of the variables obtained from the CHAIN data, are provided in green text coloring in Table 1. We recognize that the phase SI captures GNSS signal phase variations, which are potentially due to both deterministic and stochastic processes (Wang et al., 2018) and (P.T. Jayachandran, University of New Brunswick, personal communication, July 2018), but use the index in this work to represent scintillation, which is consistent with previous phase scintillation studies.

We note that though we focus on the CHAIN network here, our methods are scalable to the global complement of GNSS scintillation receivers, and such scaling is the objective of ongoing work.

Table 2
Canadian High Arctic Ionospheric Network Stations[a] *at Geographic Latitudes* >60°

	Station name (abbreviation)	Receiver model	Geographic location (degrees, [Lat, Long])	Magnetic location (degrees, [MLat, MLong])[b]
1	Arctic Bay (arc)	PolaRxS	[73.004, 274.974]	[81.078, 349.049]
2	Arviat (arv)	PolaRxS	[61.098, 265.929]	[69.956, 334.857]
3	Cambridge Bay (cbb)	GSV4004B	[69.102, 254.885]	[76.404, 313.991]
4	Coral Harbour (cor)	PolaRxS	[64.188, 276.650]	[72.948, 352.116]
5	Eureka (eur)	GSV4004B	[79.990, 274.098]	[87.265, 345.243]
6	Fort Simpson (fsi)	PolaRxS	[61.757, 238.772]	[66.959, 296.328]
7	Fort Smith (fsm)	PolaRxS	[60.026, 248.067]	[66.915, 308.847]
8	Gjoa Haven (gjo)	PolaRxS	[68.633, 264.152]	[76.824, 329.840]
9	Hall Beach (hal)	GSV4004B	[68.767, 278.744]	[77.098, 356.019]
10	Iqaluit (iqa)	GSV4004B	[63.737, 291.460]	[71.299, 15.201]
11	Kugluktuk (kug)[c]	PolaRxS	[67.818, 244.865]	[73.817, 300.305]
12	Pond Inlet (pon)	GSV4004B	[72.693, 282.045]	[80.504, 3.214]
13	Rankin Inlet (ran)	PolaRxS	[62.825, 267.885]	[71.657, 337.729]
14	Repulse Bay (rep)	PolaRxS	[66.524, 273.769]	[75.164, 347.213]
15	Resolute (res)	GSV4004B	[74.747, 264.998]	[82.357, 327.063]
16	Sachs Harbour (sac)[d]	PolaRxS	[71.991, 234.739]	[76.032, 283.310]
17	Taloyoak (tal)[d]	GSV4004B	[69.541, 266.443]	[77.774, 333.541]

Note. Only stations with the PolaRxS receiver model (listed in green) are used to avoid potential intercalibration issues with the GSV4004B receivers. Sreeja et al. (2011) found that the receiver performance is comparable. Altitude-adjusted corrected geomagnetic coordinates for 1 June 2015 (the middle of the time period of analysis in this paper) are shown. Kugluktuk data are ignored due to known bias (R. Chadwick, University of New Brunswick, personal communication, January 2018). Sachs Harbour and Taloyoak data are not available in 2015. Complete data availability between 2008 and present are provided at http://chain.physics.unb.ca/chain/pages/data_availability.

2.4. Preparation of Data for Machine Learning

In this work we choose a *supervised* machine learning model known as SVMs, discussed in detail in section 3.3. Supervised learning refers to the scenario where we know the desired output (i.e., the *correct answer*) for each data sample. Supervised learning is the guiding principle of many machine learning algorithms, including the commonly used neural network approach. Machine learning requires the data to be prepared as data *samples* or instances. For our supervised problem, each sample must contain the input data, individually known as features, and the corresponding value of the variable we intend to predict, or the *label*. Therefore, we create one sample for each CHAIN station at each observation time, in which each consists of the full set of input features detailed above in sections 2.1–2.3 and Table 1 and the corresponding predicted label (see section 3.1 for the definition of the predicted label for this study) organized as a row vector. Figures 1a and 1b provide a visual representation of the machine learning sample construction. The samples are then combined to form the machine learning database.

High-quality data are critical to any data-driven method, and we apply a robust quality control process to produce a capable database for machine learning. Our process attempts to minimize data errors and uncertainties and to remove bad data. We first discuss the CHAIN data preparation and then provide details of the solar wind and geomagnetic activity data.

First, we examine the CHAIN data. To remove observations potentially corrupted by multipath, we apply a conservative elevation mask of 30° , which is consistent with previous statistical (Prikryl et al., 2015) and machine learning-based approaches (Jiao et al., 2013) using GNSS data. Additionally, we remove all data during loss-of-lock events using the lock time recorded by the PolaRxS receiver (Septentrio, 2015). The lock time refers to the time period over which there exists an uninterrupted lock on the carrier phase signal. Cycle slips, occurring when sharp ionospheric density gradients result in loss-of-lock between the GNSS satellite and

Space Weather

10.1029/2018SW002018

Table 3
Canadian High Arctic Ionospheric Network Data Preparation Design Choices and Rationale

Variable	Details	Notes
Ionospheric pierce point	Chosen to be 110 km	Decision should reflect the phenomenon one wishes to capture. We attempt to characterize *E*-region scintillation, a critical altitude region at high latitudes (Semeter et al., 2017).
Elevation threshold	30°	Conservative choice minimizes potential multipath effects (Jiao et al., 2013).
Slant-to-vertical projection	TEC — Komjathy (1997) method S_4 and σ_Φ — Spogli et al. (2009) method	To reduce geometric effects To minimize effects of scintillation measurements made at different elevation angles (Alfonsi et al., 2011)
Interval of GNSS signal measurement computation	60 s	To minimize effects due to noise (Dierendonck et al., 1997)
Signal lock time	>200 s	To allow convergence of phase detrending filter (Mitchell et al., 2005; Sreeja, 2016)
Statistically combining simultaneous observations from multiple GNSS satellites	Median	To obtain single *super observation* to represent current scintillation conditions in the ionosphere above the receiver with robust sensitivity to outliers (McGranaghan, Mannucci, Verkhoglyadova, et al., 2017)

Note. GNSS = Global Navigation Satellite Systems; TEC = total electron content.

ground receiver (Watson et al., 2016a), and other disruptions cause interruptions to the tracking of the carrier phase signal, and the detrending filter requires time to resettle (Mitchell et al., 2005; Sreeja, 2016). We remove all data recorded during periods when lock time was less than 200 s. Finally, we reduce the data by projecting the receiver-satellite slant signals to the vertical, geolocating at the position of the receiver, and taking the median of signals from all satellites in view of a receiver at a given time. Figure 2a provides an illustration of the latter data processing step.

We use established techniques to project the slant receiver-satellite quantities (i.e., along the receiver-satellite line of sight) to the vertical. We perform projections for three quantities: (1) TEC, (2) S_4, and (3) σ_Φ. To transform slant TEC (STEC) to the vertical we follow the method of Komjathy (1997):

$$\mathrm{TEC} = \mathrm{STEC} \cdot \sqrt{1 - \cos^2(el)\frac{R_e^2}{(R_e + h_i)^2}}. \tag{1}$$

where el is the satellite elevation angle of the satellite with respect to the receiver, R_e is the radius of the Earth (6,378.137 km in this work), and h_i is the assumed height of the ionosphere. It should be noted that the value of h_i should be chosen based on the phenomenon one hopes to study. A larger assumed value of h_i will focus on the *F*-region ionosphere (altitudes are greater than roughly 150 km), where scintillation may be more closely associated with irregularities that have convected into the receiver-satellite line of sight (Watson et al., 2016a). These irregularities may have convected far from their generation location. Values of h_i in the *E* region (roughly 90–130 km), on the other hand, will target *E*-region phenomena and are primarily associated with particle precipitation (Deshpande et al., 2016). Detailed investigation of the altitude assumption for scintillation is provided by Semeter et al. (2016, 2017) and Mrak et al. (2017) and for ionospheric irregularity drift and associated scintillation effects by Su et al. (2017) and Wang et al. (2017). We note that h_i only appears in our projection of STEC to vertical, and, therefore, has relatively little impact on our results.

To project the scintillation indices to the vertical, we follow the approach identified by Spogli et al. (2009):

$$\text{vertical } S_4 = S_4(el)\sin^b(el) \tag{2}$$

for S_4 and

$$\text{vertical}\sigma_\Phi = \sigma_\Phi(el)\sin^a(el) \tag{3}$$

for σ_Φ, where a and b are chosen based on signal characteristics (see Spogli et al., 2009, for an in-depth discussion). We follow the example of Spogli et al. (2009) and use $a = 0.5$ and $b = 0.9$.

At any given time several GNSS satellites will be in view of a single receiver. To create a database commensurate with the objective of this study (i.e., to create a novel prediction method for high-latitude scintillation and establish a benchmark for future efforts) we produce a *super* observation that is the median of all simultaneous observations for a given station. Using a super observation effectively *smears* the information in space (i.e., produces a coarser resolution). Figure 2b illustrates the ionospheric spatial resolution under this approach. In the coarsest case (when the GNSS satellites are viewed at opposite sides of the sky, or at azimuth angles 180° apart, and at the minimum accepted elevation angles of 30°) and assuming an irregularity altitude of 110 km the horizontal spatial resolution becomes ~380 km or on the order of ~2–3° at high latitudes. The super observations, in general, yield finer spatial resolution, but this estimate represents an upper bound on the spatial resolution of our data, and, therefore, the capabilities of our predictive model. Note that this estimate does not take into account the motion of the receiver-satellite signal through the ionosphere over the finite collection time, though, in comparison, has a negligible impact on spatial resolution. The CHAIN science data team provides the data at a 1-min cadence. We note that the spatial and temporal scales are commensurate with the current state of the art in ionospheric prediction (Codrescu et al., 2012) and the National Oceanic and Atmospheric Administration Space Weather Prediction Center *wishlist* for ionospheric prediction (Steenburgh et al., 2014). Additionally, the predictive task we have outlined represents a balance between GNSS user needs and tractability (the challenge grows considerably to predict scintillation, for instance, for individual GNSS satellites). Therefore, we deem the approach to be aligned with our objective to establish a benchmark upon which future efforts can build and gauge success.

The final element of the data preparation pertains to the solar wind and geomagnetic activity data. The temporal relationships between these data and ionospheric scintillation are complex, including various degrees

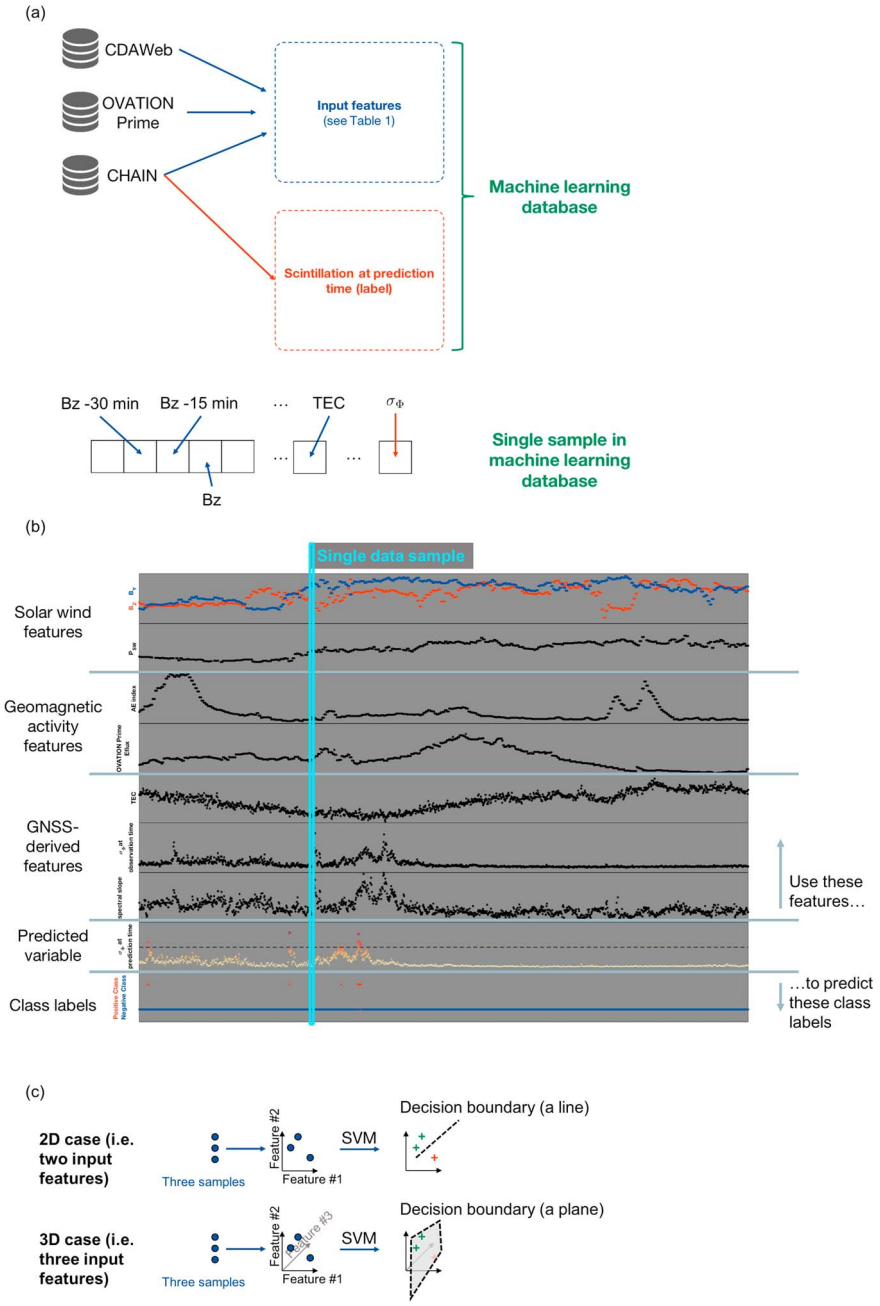

Figure 1. The critical components of the machine learning approach: (a) Construction of the machine learning database. In this work we use input from NASA's CDAWeb (https://cdaweb.sci.gsfc.nasa.gov/), the OVATION Prime (http://sourceforge.net/projects/ovation-prime/), and the CHAIN (http://chain.physics.unb.ca/chain/) to predict the phase scintillation index at a future time. (b) A generic illustration of the input features and the manner in which a data sample is constructed from time series data. The variable to be predicted is the phase scintillation index, and we formulate the prediction as a classification task in which the positive and negative classes are determined based on whether or not the phase scintillation index exceeds a given threshold (illustrated by the bottom two panels [*Predicted variable* and *Class labels*]). (c) Schematic illustration of the SVM approach to prediction (i.e., the chosen machine learning algorithm). Two easily visualized examples are shown: 2-D (top) and 3-D (bottom). For the 2-D case, the SVM algorithm finds the optimal line that separates the positive and negative prediction classes. For the 3-D case, the SVM algorithm finds the optimal plane that separates the positive and negative prediction classes. For higher-dimensional cases, an optimal hyperplane is found. The input feature space for this work contains 51 dimensions. CDAWeb = Coordinated Data Analysis Web; CHAIN = Canadian High Arctic Ionospheric Network; GNSS = Global Navigation Satellite Systems; OVATION = Oval Variation, Assessment, Tracking, Intensity, and Online Nowcasting; SVM = support vector machine; TEC = total electron content.

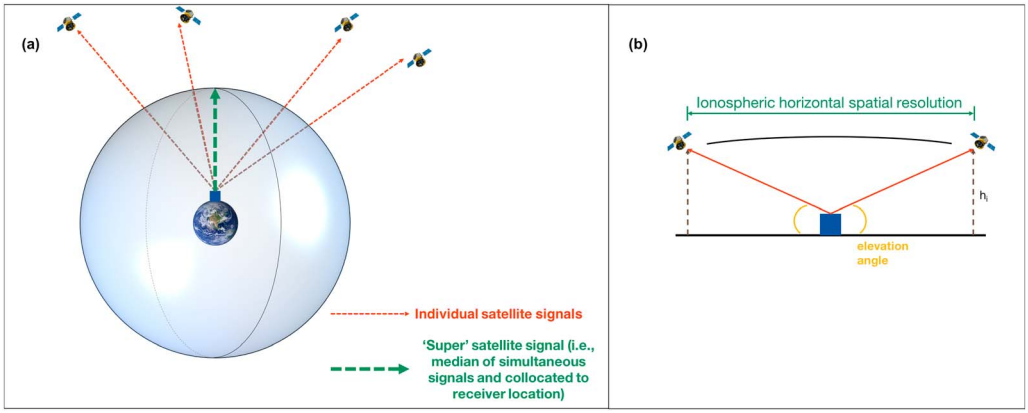

Figure 2. (a) Illustration of a single receiver simultaneously communicating with multiple Global Navigation Satellite Systems satellites (red dashed arrows) and the *super* observation created by taking the median of the data from these signals and geolocating to the receiver location (green dashed arrow). The sphere shown depicts the height of the ionosphere (h_i), which is assumed to be 110 km. (b) Schematic showing the spatial resolution as a result of the use of super observations. For the coarsest case in which the Global Navigation Satellite Systems satellites are separated by 180° in azimuth, both exist at the lowest elevation angles accepted for this work (30°), and with h_i=110 km, the spatial resolution is roughly 380 km.

of delay. We attempt to encode this information into our database in a simple yet geophysically meaningful manner. We include three data points for the solar wind and geomagnetic activity input features for each sample: (1) at the current observation time, (2) 15 min prior, and (3) 30 min prior. Precise understanding of delay times between solar wind and geomagnetic activity variations and high-latitude ionospheric phase scintillation is not well established, but studies of direct (i.e., ~0–20 min [Lu et al., 2002; Ridley et al., 1998]) and indirect (e.g., ~20–30 min [Oksavik et al., 2000]) ionospheric driving establish 0–30 min as a reasonable range to consider. We, therefore, believe that these three data points cover an appropriate range. The *Kp* and $F_{10.7}$ indices have time resolutions coarser than 1 hr and, therefore, for these variables only the value at the observation time is used. We reiterate that the solar wind data used has been propagated to the magnetopause location.

Finally, each data sample for which any input feature does not exist is discarded. Our final database consists of more than 9.6 million samples. Figure 3 shows the observational density of our database in AACGM coordinates projected onto an equal area grid to mitigate the latitudinal variation that affects fixed resolution grids at middle and high latitudes (Ruohoniemi & Baker, 1998). The equal area gridding scheme uses a constant 2° MLAT resolution and variable magnetic local time (MLT) resolution (0.28 hr at 50° MLAT to ~2.18 hr at

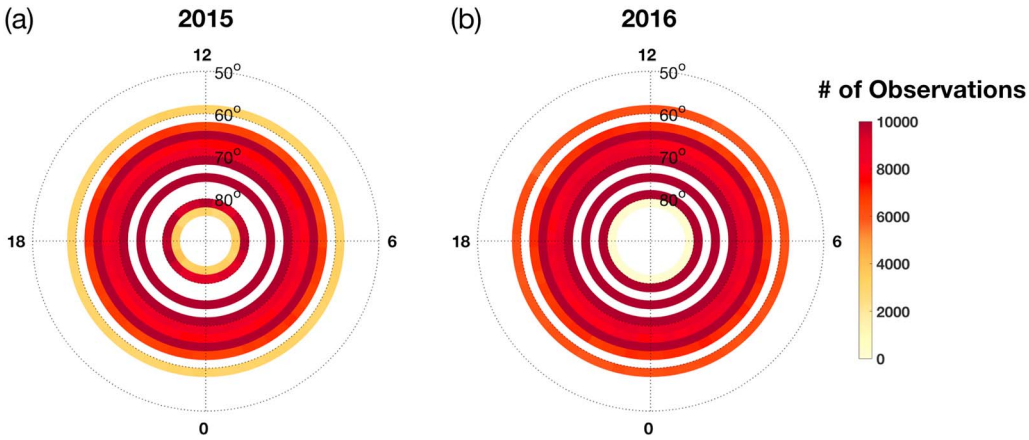

Figure 3. Observation density for the machine learning database generated from the Canadian High Arctic Ionospheric Network for (a) 2015 and (b) 2016. Data are shown on an equal area grid in altitude-adjusted corrected geomagnetic MLAT-MLT coordinates with noon MLT to the top of each polar plot and a low-latitude limit of 50° . The MLAT resolution is 2° and the MLT resolution is variable (0.28 hr at 50° MLAT to ~2.18 hr at 85° MLAT). MLAT = magnetic latitude; MLT = magnetic local time.

85° MLAT), yielding a total of 938 grid points between 50° and 90° . There is a slight difference between 2015 and 2016, due to the presence of data from the Grise Fiord station in 2015, but not 2016, and the presence of data from the Sachs Harbour station in 2016, but not 2015. There are minor gaps in latitudinal coverage in both years because of the projection and collocation process that we apply (see Figure 2 and discussion above). We obtain complete MLT coverage and all observed magnetic locations are amply sampled.

Table 3 summarily details and provides rationale for the important data processing design choices. To summarize, we take the following steps to generate the database used in this work:

1. Quality control the ionospheric GNSS data:
   (a) Apply an elevation mask of 30° .
   (b) Remove data during which the phase lock time <200 s.
   (c) Project the slant signal data to vertical and geolocate at the receiver position.
   (d) Generate a super observation as the median value among all signals acquired by the receiver at a given time (i.e., due to signals from multiple GNSS satellites in view simultaneously [see Figure 2a]).
   (e) Organize the data into samples by observation time.
2. Attach the corresponding solar wind and geomagnetic activity data.
3. Attach the corresponding predicted label (detailed next in section 3).

Figures 1a and 1b schematically show the creation of the database.

3. Prediction Methods

We apply, for the first time, a machine learning classification algorithm to the prediction of high-latitude ionospheric phase scintillation. We follow four steps to create a new machine learning prediction capability: (1) create a well-defined and explicit prediction task, (2) explore the input features, (3) select an algorithm, and (4) measure performance based on a robust evaluation metric. The following subsections, respectively, address each step.

As this paper is motivated by the value of data-driven methods, we use quantitative means to make method design choices.

3.1. Prediction Task

We investigate a quasi-predictive situation in which each input feature has zero latency (i.e., all input data are available instantaneously at a given time). Of course, in reality, various data latencies are associated with these data (e.g., GNSS and auroral precipitation); however, such considerations could be included with relatively minor difficulty in future operational circumstances.

We attempt to predict the phase SI, σ_Φ, and choose a classification task, that is, whether or not scintillation occurs based on a given σ_Φ threshold. Observations for which scintillation exceeds the threshold are called the positive class and given the label +1 and are otherwise called the negative class and given the label 0. In the machine learning community this is known as *one hot encoding*. A threshold of 0.1 rad was chosen based on two criteria: (1) a geophysically meaningful level tuned to space weather user needs (i.e., that distinguishes between conditions when scintillation is unlikely [$\sigma_\Phi \leq 0.1$] and likely [$\sigma_\Phi > 0.1$] to disrupt GNSS performance—see, for instance, Jiao et al., 2017) and (2) limits the imbalance of scintillation to non-scintillation events. To the second point, higher threshold values yield fewer scintillation events and a larger positive-to-negative class imbalance. For our database a threshold of 0.1 rad yields an imbalance ratio of ∼1:30, while a threshold of 0.5 rad grows the imbalance significantly to ∼1:1,400.

Finally, we choose a prediction interval of 1 hr; meaning, we take current data and attempt to predict whether or not scintillation will occur 1 hr in the future. Figure 1b provides an illustrative overview of the input features and predictive task. Shown are a full day of data for the CHAIN Arviat station, including an illustrative subset of the input features (solar wind, geomagnetic activity, and GNSS derived), the σ_Φ values at prediction time, and the corresponding classification label.

3.2. Exploration of Input Features

The database that we introduced in section 2 consists of 51 features. These features were selected based on the likely existence of a physical relationship with high-latitude scintillation (see, for instance, the important discriminating factors identified by Prikryl et al., 2012, for high-latitude scintillation in CHAIN data). However, higher dimensionality of input features (i.e., selecting more input features) may result in lower performance

for classification prediction tasks (Zhang et al., 2016), and performance improvement can be gained by removing less informative features. This process is typically called *featurization* and is an active area of research in the machine learning community (Khalid et al., 2014). Featurization not only results in a smaller set of input features, and, therefore, a simpler predictive model, but can also improve physical understanding by quantifying the relationships between the inputs and the predicted variable.

To perform feature selection we choose the univariate Fisher ranking score (*F* score), a widely used and computationally efficient technique. This simple approach assumes that the input features are independent and ignores correlations between them. The *F* score for feature *i* is given by (e.g., (Gu et al., 2011)):

$$F(i) = \frac{n^+ \left(\bar{x}_i^+ - \bar{x}_i\right)^2 + n^- \left(\bar{x}_i^- - \bar{x}_i\right)^2}{\frac{1}{N-2}\left[\sum_{j=1}^{n^+} \left(x_{j,i}^+ - \bar{x}_i\right)^2 \sum_{j=1}^{n^+} \left(x_{j,i}^+ - \bar{x}_i\right)^2\right]} \tag{4}$$

where $\bar{x}_i^+$ is the average of the values of feature *i* over the positive-class examples, $\bar{x}_i^-$ is the average of the values over the negative-class examples, $\bar{x}_i$ is the average of the values over the entire data set, n^+ is the total number of positives examples in the data set, n^- is the number of negative examples, and N is the total number of examples. The variance between each class for a given feature is measured by the numerator and the variance within each class for a given feature is measured by the denominator. A small *F* score (i.e., a small ratio) indicates that the two groups have similar population means while a large *F* score means the population means are distinct. Univariate feature selection relies on *F* scores to rank the input features. Our approach represents a standard statistical analysis of variance (ANOVA) univariate *F* score test, and will simply be referred to hereafter as 'univariate score'.

Figure 4 shows the univariate scores for each of the input features considered in this work (51 total). The scores contain significant geophysical information and reveal potentially important relationships. The top two *F* scores are the current value of the phase scintillation (σ_Φ) and the spectral slope, indicating that current phase scintillation conditions are more informative of the scintillation at a lead time of one hour than any other individual feature. The largest F-values could be due to several factors. We offer two explanations: 1) the lifetimes of high-latitude ionospheric irregularities in the region of the ionosphere in which a receiver-GNSS signal is disrupted are statistically greater than one hour at the scales studied here (~2-3°), or 2) the nature of geomagnetic activity (whether quiet or active) is maintained for periods longer than one hour (i.e., current conditions are indicative of conditions one hour in the future). Both explanations suggest that the ionosphere at these spatial scales exhibits 'memory' at least on a time scale of one hour and may be predictable (Siscoe & Solomon, 2006). Intuitively, the spectral slope partially describes the size of scintillation-causing irregularities (Chartier et al., 2016; Forte et al., 2016; Mezaoui et al., 2014; Wernik, 1997; Yeh & Liu, 1982) and we should expect a relationship between spatial size and/or irregularity lifetime and future scintillation (i.e., larger irregularities would be expected to affect a given area for periods exceeding that of smaller irregularities and/or have longer lifetimes).

The next four largest *F* scores are each obtained from OVATION Prime particle precipitation input features (particularly the wave and monoenergetic accelerated electrons [Newell et al., 2009]). The diffuse electrons also rank high among the input features. These findings are consistent with the conclusions drawn by Semeter et al. (2016, 2017) and Mrak et al. (2017) who identified a close connection between electron precipitation, particularly accelerated electrons, and GNSS signal corruption. A high fidelity of the OVATION Prime model would then explain its high feature score. Generally, the OVATION Prime precipitation input features appear to be more important than solar wind variables (compare, for instance, the *F* scores forOVATION Prime input features to those for $B_{y,z}$, P_{sw}, and V_{sw}). Combinations of solar wind variables (i.e., the Newell et al. [2007] and Borovsky [2013] coupling functions) and, separately, geomagnetic activity indicators (e.g., the *AE* index) contain important information for this predictive task. Finally, there are several input features with quite low univariate scores, indicating that these input features are expected to contribute little or nothing to the overall predictive capability of the SVM model. In fact, more input features may even reduce predictive performance by driving the model to overfit (Zhang et al., 2016).

It is important to note that some variables will not, by themselves, have high predictive power for scintillation, but are important in the context of other variables. We acknowledge that a more robust method to perform featurization would be a full statistical discriminant analysis of all variables and combinations of

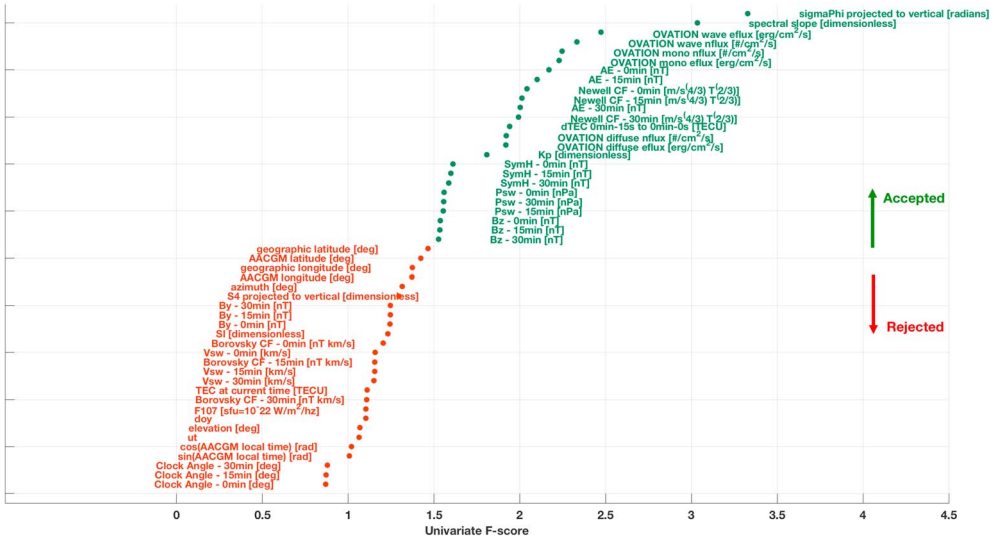

Figure 4. Univariate *F* scores for each of the 51 input features calculated for the 1-hr predictive task. The features are organized by increasing *F* score and the top 25 selected for the prediction algorithm are shown in green. AAGCM = altitude-adjusted corrected geomagnetic; AE = auroral electrojet index; CF = coupling function; OVATION = Oval Variation, Assessment, Tracking, Intensity, and Online Nowcasting; TEC = total electron content; TECU = TEC unit.

variables (i.e., only include a feature/combination of features if one can statistically reject the null hypothesis that the positive and negative class data for this feature/combination of features are drawn from the same distribution). Though we have not investigated such multivariable considerations in our feature selection analysis here (at the time of writing, no such comprehensive evaluation of variables for scintillation exists, though the authors note the value of such a study; note, for instance, the benefit of the work of Leka & Barnes, 2003, to the solar flare prediction community), we choose to include a broad range of the available input data, from which the design of the SVM algorithm inherently considers such relationships in searching for the optimal decision boundary. This responds to the intuitive expectation that, for a complex variable such as scintillation, no single variable is capable of distinguishing between scintillation and non-scintillation populations (and, in fact, even low-order combinations of variables such as three or four are still unlikely to be able to provide sufficient discrimination), but that multivariable combinations are more effective at distinguishing between populations. Numerous studies have reached a similar conclusion, notably Gjerloev et al. (2018), indicating an important theme for ionospheric specification and prediction. Given the extent of the input features we consider, we believe our machine learning database is robust and capable for this prediction exploration.

To examine the impact of the number of input features on the SVM model performance, we perform a sensitivity analysis in which the number of input features is varied from 1 to 51 (i.e., the total number of input features collected for this study; see Table 1 and section 2) in the order of decreasing univariate score as shown in Figure 4. In other words, we begin with a single feature, the current value of σ_Φ. Then, we add the next most informative feature one at a time, examining the resultant SVM performance at each step, until we have included the complete set of input features. At each step we train 10 SVM models using different subsets of 50K training data samples, testing the performance on the remaining data, and record the average and standard deviation of various evaluation metrics. The results revealed little variation between SVM models across training/testing subsets for a set number of input features and were relatively invariant to the number of models. Therefore, we deemed 10 to be sufficient. Figure 5 shows the outcomes of this sensitivity analysis for the TSS (see section 3.4 for the definition of TSS) evaluation metric and for the first 30 input features. The blue and red scatter points respectively show the average and standard deviation error bars of the TSS computed for the 10 SVM models at each step, and the *x* axis reveals the input feature that was successively added beginning with the first (*sigmaPhi projected to vertical* [*radians*]) on the far left. The average and standard deviation error bars for the persistence case (i.e., assuming the phase SI remains the same for the next hour) are shown by the horizontal solid and dashed black lines, respectively, and are quantitatively 0.41 $\pm$ 0.01. The dashed green vertical line marks the point beyond which input features are no longer included in the SVM and were used to produce the model results presented throughout this paper (i.e., 25 input features).

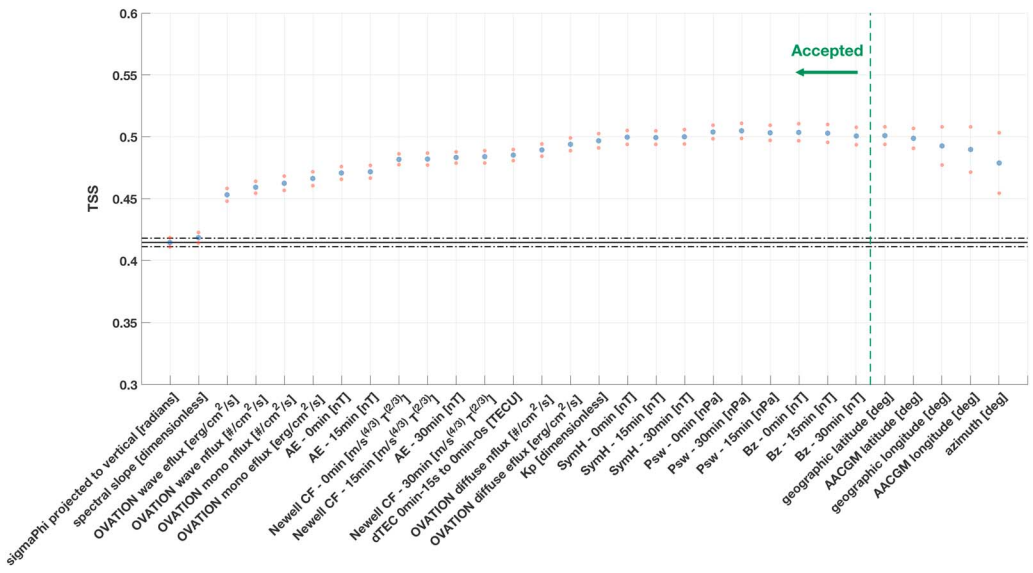

Figure 5. TSS versus number of input features used to train a Support Vector Machine (SVM) model. The blue and red scatter points, respectively, show the average and standard deviation error bars of the TSS for the ten SVM models generated for each input feature set. Moving left to right the x-axis reveals the input feature that was successively added. The average and standard deviation error bars for the persistence case are shown by the horizontal solid and dashed black lines, respectively, and are quantitatively 0.41±0.01. The dashed green vertical line marks the point beyond which input features are no longer included in the SVM model results presented throughout this paper (i.e., 25 input features). Note that the vertical axis limits are 0.3-0.6 to improve clarity. AACGM = altitude-adjusted corrected geomagnetic; AE = auroral electrojet index; CF = coupling function; dTEC = differential TEC; OVATION = Oval Variation, Assessment, Tracking, Intensity, and Online Nowcasting; TSS = true skill score.

Figure 5 reveals several important points. First, additional input features increase the predictive capability of the SVM model as evaluated by the TSS, however only to a certain point. Beyond the first 25 input features (marked by the dashed green line), the TSS starts to decrease and the variation of the models increases; one indication of overfitting. We, therefore, select only the top 25 input features according to univariate score. These features are shown in green in Figure 4. Figure 5 also shows that much of the skill of the SVM model for the 1-hr predictive task is obtained by the current scintillation conditions (i.e., current value of σ_Φ and the spectral slope). The importance of the current conditions will obviously decrease as prediction time is increased, which we address in section 5.3 and will more thoroughly investigate in future work.

Finally, given the drastically different dynamic ranges of the input features, each feature is normalized to a zero mean unit variance, which is a common and often necessary data preparation step prior to training a machine learning model (Ng, 2018; Witten & Frank, 2005). The scikit-learn package *RobustScaler* function is used to normalize the data while remaining robust to outliers, meaning that the median is removed and the data are scaled according to the interquartile range (25th to 75th percentiles).

3.3. SVM for Classification

The SVM (Cortes & Vapnik, 1995) is a popular classification algorithm due to its high accuracy, capability to handle high dimensional data, and flexibility to model diverse data (Schölkopf et al., 2004). The mathematical formalism of the SVM is well addressed in the literature (e.g., Burges, 1998; Hastie et al., 2001), including in geospace applications (Jiao et al., 2017), and we only give brief introduction to it here, focusing instead on why it is an appropriate choice to the problem at hand.

Our problem is given an input data sample, $\mathbf{x}_m \in \mathbb{R}^N$ where N is the number of input features that defines the problem dimensionality, and the corresponding classification label, y_m (either +1 or 0 in our two-class scintillation application), find a function that separates the data based on their associated label. Note that each label corresponds to a *class*, which in this work is either scintillation or no scintillation. If the classes are linearly separable in the feature space, then this function is a hyperplane of the form $f(\mathbf{x}) = \mathbf{x}^\intercal\beta + \beta_0$ such that $\mathbf{x}_m$ is of class $y_m = +1$ if $f(\mathbf{x}) \geq 0$ and class $y_m = 0$ if $f(\mathbf{x}) < 0$. The SVM operates by maximizing the distance, or margin, between the decision hyperplane and the closest data samples. These samples are called

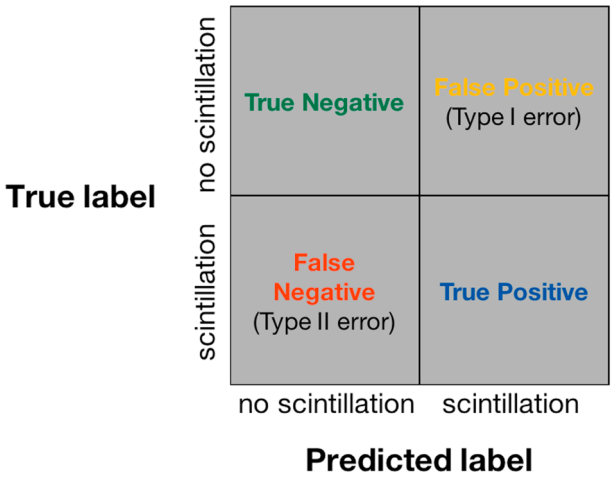

Figure 6. The contingency matrix. On the left axis are the true labels for the predicted variable (i.e., scintillation or no scintillation) and on bottom axis are the predicted labels. This matrix defines four critical categories: True Negative for which no scintillation occurs and no scintillation is predicted, True Positive for which scintillation occurs and scintillation is predicted, False Negative for which scintillation occurs yet no scintillation is predicted, and False Positive for which no scintillation occurs yet scintillation is predicted. These four values provide the basis from which all prediction metrics are derived.

support vectors, lie closest to the decision boundary, and are the most difficult to classify. The support vectors determine the hyperplane. Figure 1c shows illustrations of the SVM approach for two- and three-dimensional cases, each containing three training data samples for illustrative purposes.

However, our input data samples are likely not linearly separable in feature space ($\mathbb{R}^N$). A common solution is to map the data samples nonlinearly to a higher-dimensional space using a kernel basis function $\Phi_k(\mathbf{x})$, $k = 1, \ldots, K$ and to identify the discriminating hyperplane in this higher-dimensional space, $\hat{f}(\mathbf{x}) = \Phi(\mathbf{x})^\intercal \hat{\beta} + \hat{\beta}_0$. In the SVM literature this is called the *kernel trick* and its use to produce a more capable classifier derives from Cover's theorem, which states that complex, linearly inseparable classification problems are more likely to be linearly separable when cast nonlinearly (via the kernel function) into much higher-dimensional spaces (Cover, 1965). Therefore, the classification is now nonlinear in the original feature space $\mathbb{R}^N$. A common choice for the kernel function, and that chosen for this work, is the Gaussian kernel or radial basis function: $k(\mathbf{x}, \mathbf{x}') = e^{-\gamma|\mathbf{x}-\mathbf{x}'|^2}$, where γ is the width parameter, effectively determining the influence of a single data sample, $\mathbf{x}$. When γ is small all samples affect the decision boundary, and a smooth boundary is identified, but as γ is increased the affect of each sample becomes more local. If γ is set too large, then, overfitting can occur and it is, thus, an important design parameter.

The formulation above requires the correct classification of all samples, however improved performance can be obtained by allowing some data samples to be misclassified, especially with noisy data samples. This is accomplished through regularization where a term C determines the penalty for misclassified samples or those that exist in the margin. The larger the value of C, the heavier the penalty for misclassification and margin errors. This regularized version of SVM is called a *soft margin classifier*.

Class imbalance is a major problem in many machine learning algorithms. Many algorithms are prone to strong bias toward the majority class, in this case non-scintillation, at the cost of neglecting the minority one. There are several approaches to mitigating this problem (e.g., Longadge & Dongre, 2013), but a particularly effective technique is to separately set C for the negative (majority) and positive (minority) classes. We explore this approach in our results, introducing the *class weight ratio* ($w_{\text{negative class}} : w_{\text{positive class}}$) as another design parameter.

Together, γ, C, and class weight ratio are called *hyperparameters* of the SVM and their selection, or tuning, is critical to predictive performance. We implement the SVM approach using the python scikit-learn library (Pedregosa et al., 2011), which is open source, widely used, and well established. Additionally, we openly and freely provide sample software to produce the results shown here through a FigShare Project with the same name as this paper and available at https://doi.org/10.6084/m9.figshare.6813143 (McGranaghan et al., 2018a). Data from this work are also provided through the FigShare Project at https://doi.org/10.6084/m9.figshare.6813131 (McGranaghan et al., 2018b).

3.4. Evaluation

There are numerous metrics through which the success of a prediction method can be evaluated and each are derived from four values: false positives (FPs), false negatives (FNs), true positives (TPs), and true negatives (TNs). These values are defined by the contingency table. Figure 6 shows the format of the contingency matrix and defines the terms. The textual definitions of these categories are as follows: True Negative (TN) for which no scintillation above the threshold occurs and no scintillation is predicted, True Positive (TP) for which scintillation occurs and scintillation is predicted, False Negative (FN) for which scintillation occurs yet no scintillation is predicted, and False Positive (FP) for which no scintillation occurs yet scintillation is predicted. These four values provide the basis from which all prediction metrics are derived. Robust quantitative measures are essential to compare the performance of different prediction methods. Table 4 details common evaluation metrics and those which guide our investigation of SVMs, though we primarily focus on the TSS in this paper.

Table 4
Model Evaluation Metrics

Metric	Equation	Significance/value	Shortcoming
Precision	$\frac{TP}{TP+FP}$	Capability of the model to identify only	Sensitive to class imbalance
		the positive cases	
Recall	$\frac{TP}{TP+FN}$	Capability of the model to identify all of the positive	…
		cases; insensitive to class imbalance	…
Specificity	$\frac{TN}{TN+FP}$	Capability of the model to identify the negative cases	Sensitive to class imbalance
*F*1	$2\frac{Precision \times Recall}{Precision+Recall}$	Harmonic average of precision and recall; ranges over [0:1] such that $F1 = 1$	…
		indicates perfect precision and recall	…
Area under precision-recall	—	Better indication of model performance	…
curve (AU-PRC)		with large class imbalances	…
True skill score (TSS)	$\frac{TP}{TP+FN} - \frac{FP}{FP+TN}$	Accounts for random chance and is insensitive to class imbalance	…
		Ranges over [−1:1] such that $TSS = -1$ means every event is incorrectly	…
		classified; $TSS = +1$ means every event is correctly classified; and $TSS = 0$	…
Bloomfield et al. (2012)		means the model predicts consistently with a random chance predictor	…

Note. FN = False Negative; FP = False Positive; TN = True Negative; TP = True Positive.

Given the findings of Bloomfield et al. (2012) we adopt the TSS as the primary metric by which to judge model prediction capability. We suggest, based on the body of work from various communities (notably weather forecasting [Manzato, 2005] and solar flare forecasting [Barnes et al., 2016; Bloomfield et al., 2012; Bobra & Couvidat, 2015; Jonas et al., 2017]) as well as our own exploration detailed in this manuscript, that TSS is a robust and appropriate measure to benchmark high-latitude ionospheric phase scintillation prediction, a key objective of this work. To our knowledge this paper is the first to compute a TSS from a forecastable model of high-latitude phase scintillation and to use this metric to evaluate prediction models.

It should be noted that one potential challenge of using TSS to evaluate success is that it treats FPs (i.e., false prediction alarms) and FNs (i.e., missed occurrences) in the same way and does not consider the different consequences associated with each. In the case of ionospheric scintillation it may be more costly to miss an occurrence than to falsely predict one will occur. For instance, in the case of communication the cost of a false prediction alarm may be delaying the action associated with that communication until after the prediction alarm concludes, whereas the cost of missing the occurrence of scintillation is an incomplete, corrupted, or lost communication and the resultant, potentially dangerous, disconnect between the sending and receiving parties. Clearly, the relative importance of the various entries of the contingency matrix is application specific.

4. Inherent Limitations

As important as understanding the capabilities of a prediction model is understanding where it is incapable. All models will have shortcomings, and their effective use depends on their clear identification. Therefore, we list the limitations that accompany this work:

- The input data are not fully representative of the vast spectrum of circumstances that can be manifested in the ionosphere, and our prediction model is only capable of robustly predicting circumstances that are represented in the input data (2015–2016). Predictions outside of this range can be made, and, in general, may be accurate, but are, however, extrapolation.
- Related to the first point, the input data cover only the declining phase of solar cycle 24, and the same point regarding predictions for different ranges of input data applies for predictions during different phases of the solar cycle.
- We use only the CHAIN receiver network, which does not geographically cover all latitudes and longitudes. It does, however, provide complete coverage of the high-latitude ionosphere in magnetic coordinates.
- We use super observations (see section 2.4) which is a statistical summary of the ground receiver-to-individual GNSS satellite links available at a given time for a given station. Statistically summarizing the data in this manner inevitably reduces the information content of the original data (primarily by

reducing the spatial resolution of the data), which may preclude our results from impacting certain scientific investigation (e.g., at resolutions finer than those provided by the super observations).
- The predictive task we have defined may not be appropriate to all applications. We address this limitation briefly in section 5.3 by examining alternative tasks. Full exploration of these tasks will be the focus of future publication.
- We explore only one machine learning algorithm, the SVM.

Finally, it is important to be aware of the resolution and data preparation aspects to understand the model capabilities. We have detailed these items in section 2.

5. Results

We reiterate here that the objective of this paper is to use machine learning algorithms to create a novel prediction method for high-latitude phase scintillation and to establish a benchmark for future efforts. Therefore, we separate our results into two sections: (1) exploration of SVM machine learning models (section 5.1) and (2) presentation of the capabilities of the SVM approach (section 5.2). We primarily use the TSS to evaluate predictive performance in order to allow comparison with future studies (Bloomfield et al., 2012). Throughout the results section we use data from 2015 for training. Section 5.2 presents a validation period, taken from 2016, to assess the model performance. Throughout the results the model is always evaluated based on predictions for data that were *not* used in training (i.e., independent data samples).

5.1. SVM Model Exploration

The number of parameters and design choices that influence the construction of an SVM model create a vast exploration space. Like any machine learning or physics-based modeling attempt, it is important to gain an understanding of the sensitivity of the model on the input parameters. Therefore, we attempt a robust, quantitative analysis of this exploration space. Our objective is to identify the set of (1) SVM hyperparameters and (2) training data size that collectively yield high predictive performance. To this end, we organize our SVM model exploration through a series of important questions for SVM model design. These questions not only guide the exploration process for these machine learning models but also provide a road map for future efforts to develop high-latitude scintillation prediction models.

5.1.1. What Effect Do the Penalty Parameter (C), the Width of the Gaussian Kernel Basis Function (γ), and the Class Weight Ratio Have on the SVM Performance?

The three primary design parameters, or hyperparameters, for an SVM model are the penalty parameter (C), the width of the Gaussian kernel basis function (γ), and the class weight ratio. We designed a sensitivity analysis whereby a range of each of the hyperparameters was defined and a separate SVM model was trained for each point in the search space. The ranges used were $C = \{0.001, 0.01, 0.1, 10\}$; $\gamma = \{0.001, 0.01, 0.1, 10\}$; and class weight ratio = {1:1, 1:2, 1:10, 1:50, 1:100}. Therefore, a total of 125 SVM models were examined. For computational reasons discussed further below in section 5.1.2 we used a subset of 50K data samples from 2015 to train each SVM model and computed the TSS based on the remaining data samples in 2015 (~4.4 million). This may seem like a small number of samples for training; however, recall that the SVM model identifies the small number of most important data samples (i.e., the *support vectors*) and determines the decision hyperplane based on them. We acknowledge that a limited data set provides fewer support vectors from which to determine the discriminating hyperplane, but there is evidence that SVM can obtain high accuracy with a small fractional training data set (Mourad et al., 2017). We explore this quantitatively through a sensitivity analysis based on the number of data samples used for training in section 5.1.2 below. Future work will address intelligent subsampling of the available data to attempt to identify the support vectors a priori to optimally increase model predictive capability and minimize training time. Figures 7a to 7e show the outcome of the sensitivity analysis as a function of TSS for increasing class weight ratio (i.e., moving left to right, greater weight is progressively placed on correctly predicting the positive class).

Figures 7a and 7b show that at equal or low class weight ratios, relatively independent of the values of C and γ, low skill scores are obtained. Skill increases noticeably for class weight ratios greater than 1:2 and peaks in this coarse grid search for a ratio of 1:50. Figures 7c and 7d suggest that SVM models with predictive skill exist in the range of class weight ratios between 1:10 and 1:100 and primarily for C values between 0.01 and 10 and γ values between 0.01 and 0.1.

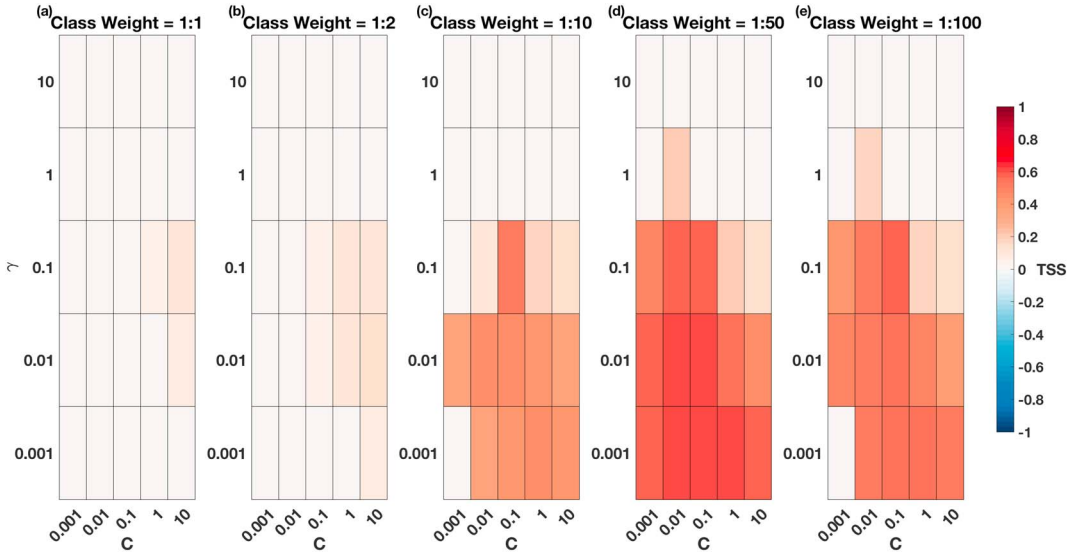

Figure 7. Support vector machine hyperparameter sensitivity analysis. For each combination of hyperparameters from the ranges: $C = \{0.001, 0.01, 0.1, 10\}$; $\gamma = \{0.001, 0.01, 0.1, 10\}$; and class weight ratio = {1:1, 1:2, 1:10, 1:50, 1:100}, a support vector machine model was trained using 50K data samples from 2015 and the TSS computed from the remaining 2015 samples (~4.4M). (a–e) The results as a function of increasing class weight ratio (i.e., greater weight placed on correctly predicting the positive class). TSS = true skill score.

5.1.2. What Effect Does the Training Data Size Have on the SVM Performance?

SVMs determine the optimal decision boundary by identifying the support vectors, a small subset of data samples that most influence the boundary.

The computational cost of training an SVM model grows nonlinearly with respect to the number of training samples (the exact scaling is dependent on a number of factors, including the number of support vectors [and, therefore, the hyperparameters C and γ] and the specific algorithm used, and is generally between $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$ in computational cost; see Chapelle, 2007, and Chapter 12 Sections 2 and 3 of Hastie et al., 2001). Therefore, practically speaking, computational considerations place a limit on the number of training samples that can be used. Therefore, it is important to investigate the effect of training sample size on SVM performance.

Based on the results shown in Figure 7 we use the following hyperparameters to assess the impact of training data size on predictive performance: $\gamma = 0.1$, $C = 0.1$, class weight = 1:10. We carried out a sensitivity study whereby SVM models were trained using different input training data sizes in the range: {50K, 100K, 250K, 1M}. For each training data size, we train 50 separate SVM models each using a different random subset of training data samples and show the average value (red dots) and standard deviation (blue error bars) across the models in Figure 8a. We also present the growth in training time in Figure 8b.

Figure 8a shows that the TSS approaches a plateau for a given SVM model as training data size is increased, indicating that subsets of training data can capably be used to train an SVM model. Additionally, the variation from using different subsets of training data decreases with increasing training data size (smaller error bars on the 1M case compared with the 50K case in Figure 8a) but is relatively small for each of the data sizes tested. Figure 8b shows the drastic growth in training computation time required as the training data sample size is increased, growing from the order of hours for thousands of samples to the order of weeks for millions of observations. This clearly illustrates the need to balance number of training data samples and training time.

5.2. Exploration of SVM Model Capability

In section 5.1 we comprehensively examined the SVM design space for the 2015–2016 CHAIN machine learning database and the 1-hr predictive task and evaluated the models using the TSS. Here we use the outcomes of that exploration to examine a candidate SVM model (100K training data samples selected randomly from 2015, $C = 0.1$, $\gamma = 0.01$, and class weight ratio = 1:50), applying predictions to a particular period from 15–25 January 2016. This allows us to test the model on data unused during the training and testing periods and to incorporate space weather understanding to interpret, understand, and, ultimately, improve the predictive model. It is important to note that this study does not claim to identify an optimized SVM model

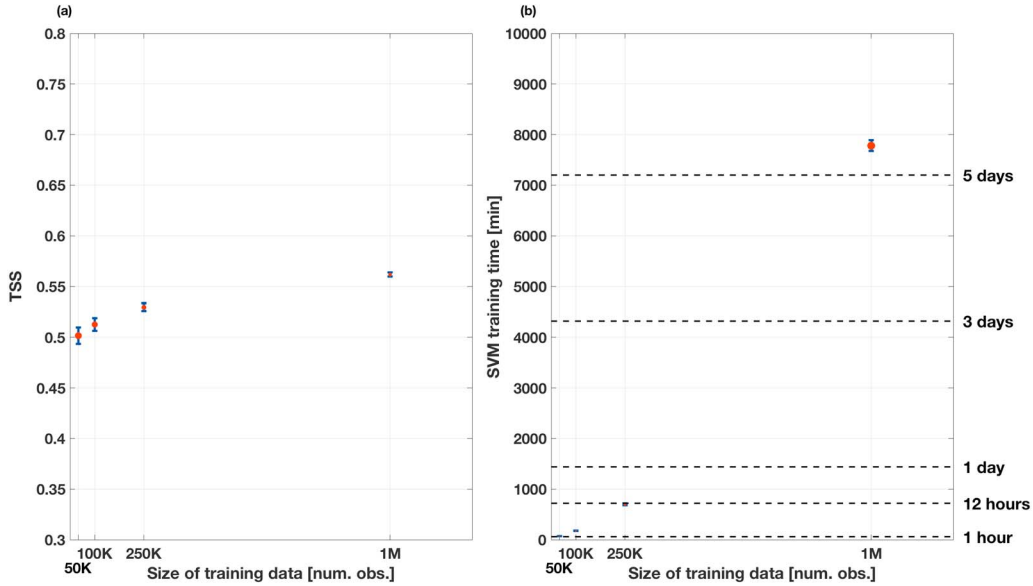

Figure 8. SVM input training data size sensitivity analysis. (a) TSS and (b) SVM training time across the range of training data sample sizes tested: 50K, 100K, 250K, and 1M. For each training data size, we train 50 separate SVM models each using a different random subset of training data samples and show the average value (red dots) and standard deviation (blue error bars) across the models. SVM = support vector machine; TSS = true skill score.

for high-latitude scintillation prediction but rather selects an SVM model based on our findings above and assesses it during a specific period, which allows us to examine SVM performance in the context of space weather knowledge (McGranaghan, Bhatt, et al., 2017).

Figure 9 presents a case study during this period on 20 January 2016 for a single station, McMurdo, taken to be representative and illustrative of several important points. Figures 9a to 9c provide contextual solar wind and geomagnetic activity data: (a) the solar wind IMF B_z and B_y components (nT), (b) the planetary *Kp* index, and (c) the *AE* index (nT). Figures 9d to 9f show the phase SI (σ_Φ) values (the true values of σ_Φ are repeated in each of Figures 9d to 9f), where the values in Figures 9e and 9f have been color coded according to the contingency matrix entries (shown to the left of the figures) of the SVM and persistence predictions, respectively.

The IMF B_z component shows a southward turning of the magnetic field followed by a prolonged period of southward-directed magnetic field that corresponds to enhancements in geomagnetic activity as indicated by the *Kp* and *AE* indices. McMurdo station, located at 63° magnetic latitude observes phase scintillation enhancements, primarily around 1000–1500 UT and throughout the postmidnight local time sector. This behavior could be indicative of substorm-induced scintillation. Referring to the persistence predictions in Figure 9f it is clear that persistence is incapable of capturing this scintillation, yielding many FNs shown by red points and very few TPs shown in green. Alternatively, the SVM model capably predicts the scintillation in the postmidnight local time sector, correctly predicting the most extreme scintillation as indicated by the TPs in green. However, the characteristic of the SVM model predictions to overpredict during active times is also clear. Throughout the active period, and for a prolonged period after the return of the phase scintillation to values below the 0.1-rad threshold (shown by a dashed black line in Figure 9d), the SVM model yields FPs (yellow points). These results, and other case studies that were analyzed, illustrate that the SVM model over-predicts scintillation when the phase SI approaches the threshold. However, for the prediction of scintillation it is more costly to miss important scintillation events such as that around 1200 UT in Figure 9 than to incorrectly predict cases of no scintillation. It is quite likely that attempting to predict when a single data point for a highly dynamic variable will exceed a given threshold causes the performance issues for the SVM model, and we address this shortcoming in section 5.3.

We can obtain further information about the capability of the SVM model by examining predictions from all stations throughout the 15–25 January 2016 period as a function of AACGM latitude and local time. Figure 10 shows the number of predictions on polar plots laid out according to the contingency matrix (e.g., TNs in the upper left quadrant and TPs in the bottom right quadrant). Each polar plot is displayed in the same way. Data

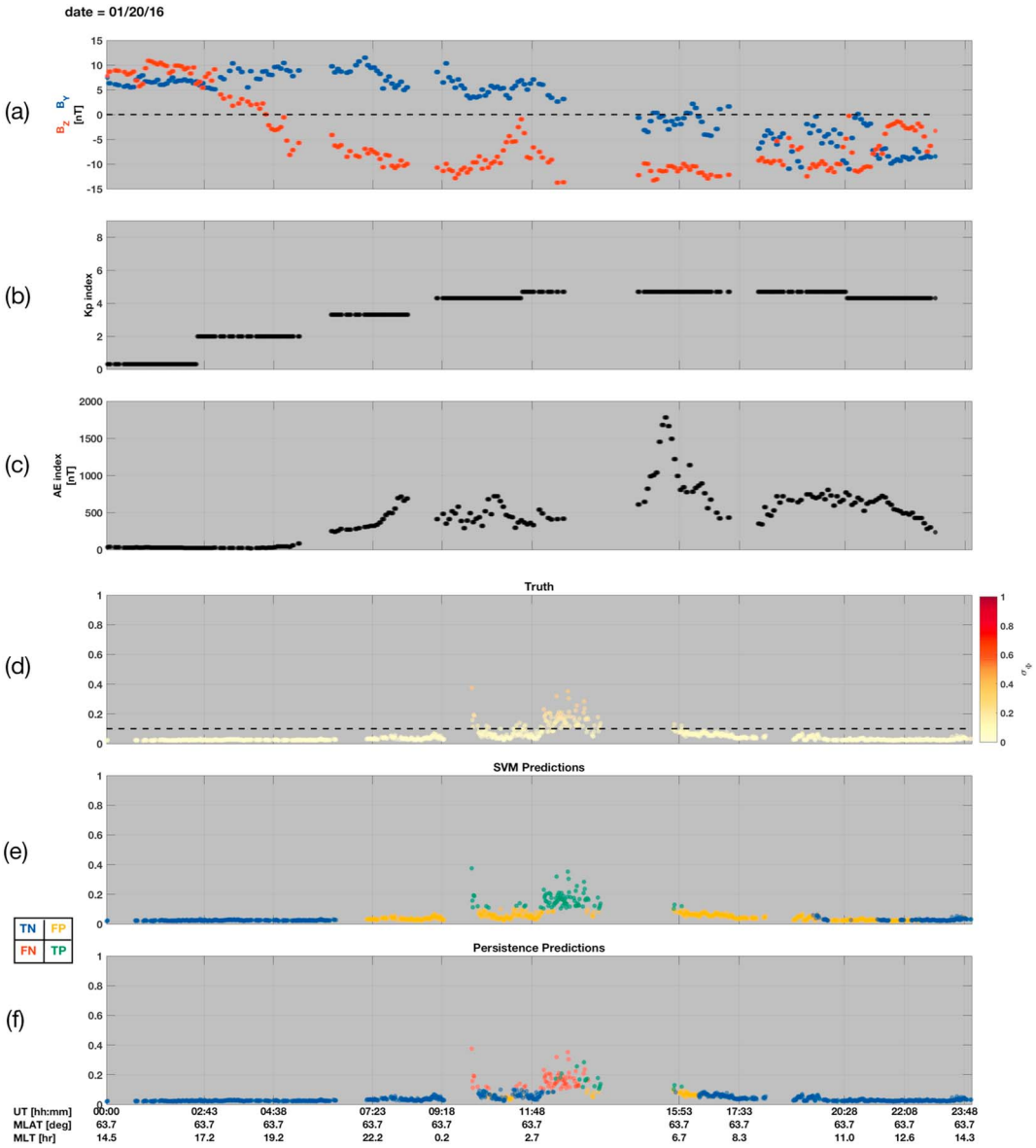

Figure 9. The 20 January 2016 case study. Shown are contextual data: (a) solar wind interplanetary magnetic field B_z and B_y components (nT), (b) the planetary *Kp* index, and (c) the *AE* index (nT). The bottom three panels show (d) true phase scintillation index values, (e) SVM predictions, and (e) persistence predictions. All data are shown at the time of prediction. Phase scintillation index data in panels e and f are the same as the time series shown in panel d (the true σ_Φ values) but have been color coded according to the contingency matrix shown to the left of the panels based on the scintillation/no scintillation classification predictions of these models. The 0.1-radian threshold used for this work is shown by a dashed black line in panel d. Data for the McMurdo ground station are selected as a representative case and the UT, altitude-adjusted corrected geomagnetic latitude and local time of the station are provided as *x*-axis labels. AE = auroral electrojet; MLAT = magnetic latitude; MLT = magnetic local time; SVM = support vector machine; UT = universal time; FN = False Negative; FP = False Positive; TN = True Negative; TP = True Positive.

are shown on an equal area grid in AACGM MLAT-MLT coordinates with noon MLT to the top of each polar plot and a low-latitude limit of 50° . We use the same equal area binning scheme in Figure 10 as that used in Figure 3 and described above. Note that the color ranges are different for the no scintillation cases (top row of both Figures 10a and 10b) than for the scintillation cases (bottom rows) due to the difference in number of occurrences.

Comparing the bottom right of Figures 10a and 10b (the TPs) we find that the SVM model captures more of the scintillation occurrences throughout the polar and auroral regions, and the improved performance

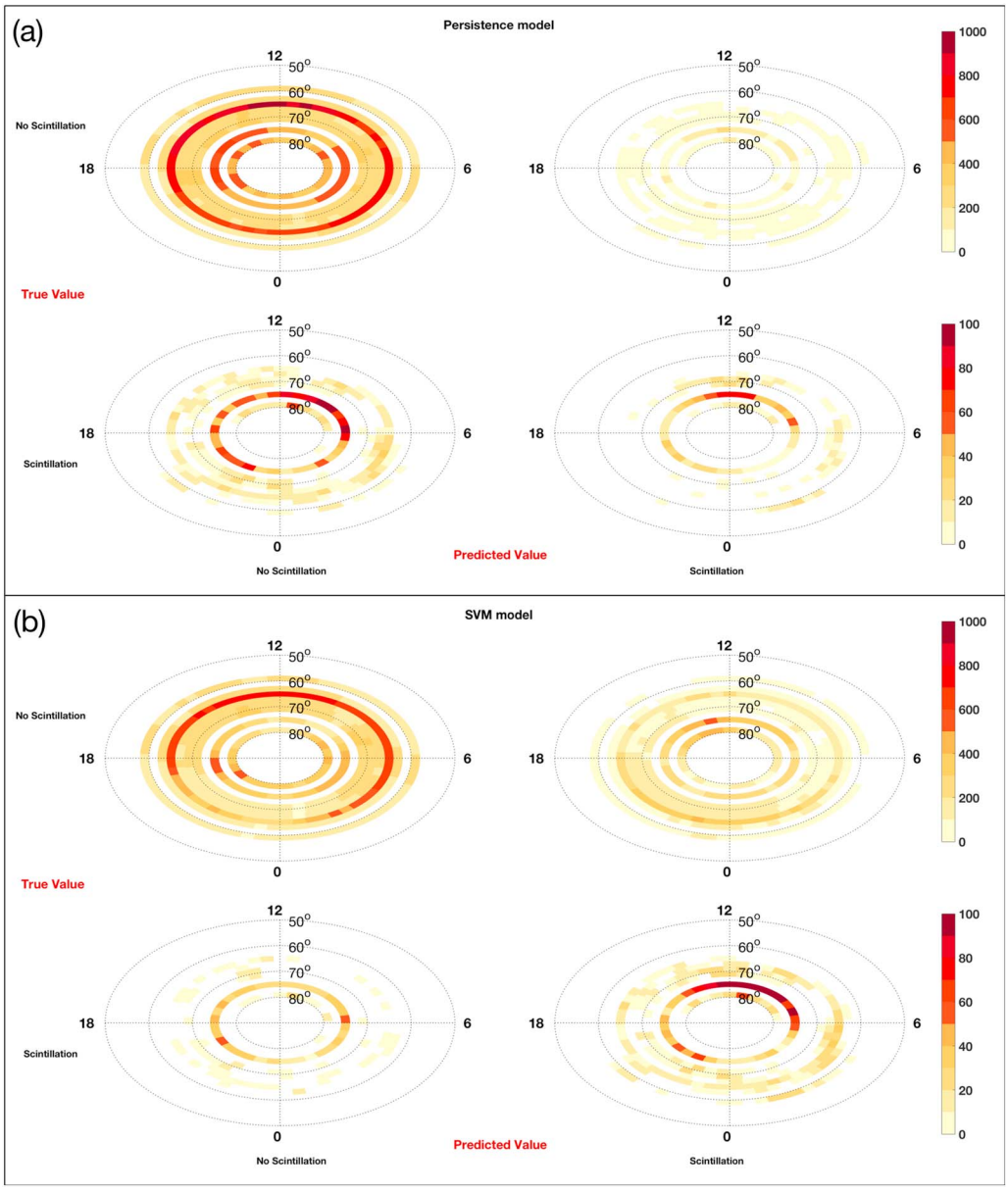

Figure 10. Polar plots showing the (a) persistence and (b) SVM model predictions during the 15–25 January 2016 period laid out according to the contingency matrix entries. Note that the color ranges are different for the no scintillation cases (top row of both a and b) than for the scintillation cases (bottom rows) due to the difference in number of occurrences. Each polar plot is displayed in the same way. Data are shown on an equal area grid in altitude-adjusted corrected geomagnetic MLAT-MLT coordinates with noon MLT to the top of each polar plot and a low-latitude limit of 50° . The MLAT resolution is 2° , and the MLT resolution is variable (0.28 hr at 50° MLAT to ~2.18 hr at 85° MLAT). MLAT = magnetic latitude; MLT = magnetic local time; SVM = support vector machine.

over persistence is particularly pronounced in two critical areas: (1) the polar cusp (roughly a region spanning 1–2° MLAT and near-1200 MLT [Newell & Meng, 1988]) and (2) the nightside premidnight to postmidnight auroral region. The SVM predictions exhibit both a higher number of TPs and a lower number of FNs in these regions. Consistent with the findings shown in Figure 9, however, the SVM model does exhibit larger numbers of FPs. The amount of FPs (i.e., the degree of overprediction) is influenced by the class weight ratio chosen and can, therefore, be optimized using the SVM approach. Finally, there is no discernible pattern in the TNs for either predictive model, indicating that these conditions are ubiquitous and both models capture these cases independently of MLAT-MLT location. The TSSs for this period for the persistence and SVM models are 0.25 and 0.48, respectively.

Table 5
Prediction Tasks Evaluation and Definitions

Task	FN	FP	TN	TP	TSS	Definition of scintillation occurrence
SVM: 1 hr, 1 point	1,071	32,188	89,718	3,252	0.49	If **a single** σ_Φ point exceeds the threshold at a prediction time of **1 hr**
SVM: 1 hr, 3 points	638	18,218	106,156	1,217	0.51	If **3 consecutive** σ_Φ points exceed the threshold at a prediction time of **1 hr**
SVM: 1 hr, 6 points	267	7,934	117,626	402	0.54	If **6 consecutive** σ_Φ points exceed the threshold at a prediction time of **1 hr**
SVM: 1 hr, 15-point mean	1,053	29,337	92,728	3,111	0.51	If **the average of 15 consecutive** σ_Φ points exceeds the threshold at a prediction time of **1 hr**
SVM: 3 hr, 1 point	940	37,865	83,961	3,463	0.48	If **a single** σ_Φ point exceeds the threshold at a prediction time of **3 hr**
SVM: 3 hr, 3 points	824	16,865	107,316	1,224	0.46	If **3 consecutive** σ_Φ points exceed the threshold at a prediction time of **3 hr**
SVM: 3 hr, 6 points	434	8,376	117,055	364	0.39	If **6 consecutive** σ_Φ points exceed the threshold at a prediction time of **3 hr**
SVM: 3 hr, 15-point mean	879	34,193	87,619	3,538	0.52	If **the average of 15 consecutive** σ_Φ points exceeds the threshold at a prediction time of **3 hr**
Persistence: 1 hr, 1 point	3,131	3,339	118,567	1,192	0.25	—
Persistence: 1 hr, 3 points	1,252	3,928	120,446	603	0.29	—
Persistence: 1 hr, 6 points	416	4,278	121,282	253	0.34	—
Persistence: 1 hr, 15-point mean	2,915	3,282	118,783	1,249	0.27	—
Persistence: 3 hr, 1 point	3,667	3,795	118,031	736	0.14	—
Persistence: 3 hr, 3 points	1,724	4,207	119,974	324	0.12	—
Persistence: 3 hour, 6 points	678	4,411	121,020	120	0.12	—
Persistence: 3 hr, 15-point mean	3,651	3,765	118,047	766	0.14	—

Note. FN = False Negative; FP = False Positive; TN = True Negative; TP = True Positive. The bold emphasis on some of the entries in this table is used to emphasize the difference between row entries, which are similar but vary in number of data points and prediction time.

The general conclusion from Figure 10 is that the SVM model more capably identifies scintillation occurrence than persistence (75% versus 28% TP rate for this time period), which corresponds also to a much lower FN rate (25%) for the SVM model than the persistence model (72%). The SVM model, however, does exhibit a tendency to overpredict (~27% FP rate), which can be mitigated through further optimization of the class weight ratio hyperparameter. Given the importance of prediction *hits* (i.e., TP identification), we believe these results indicate a significant predictive improvement using the SVM model.

5.3. Examination of Different Predictive Tasks

We recognize that the 1-hr, single data point predictive task, while well-defined, important, and conducive to our objective of establishing a benchmark for high-latitude ionospheric phase scintillation, does not encapsulate the full prediction space of interest. Therefore, we examine an extended set of predictive tasks briefly here. We focus on three aspects of the prediction problem: (1) extended lead times, (2) event-based prediction, and (3) prediction of time-averaged phase scintillation. For all tasks we use a standard SVM model design chosen based on the results shown in Figure 7 and consistent with the model used to produce Figures 9 and 10 (100K training data samples, $C = 0.1$, $\gamma = 0.01$, class weight = 1:50). We evaluated each of the predictive tasks during the same ten day period between 15 and 25 January 2016 addressed in section 5.2. We reiterate that this period represents novel data (i.e., unique from the data used for training) and shows a range of geomagnetic activity in order to provide information about performance across a large range of geospace conditions, including both quiet and active periods. We created separate SVM models for eight different predictive tasks. Table 5 provides a description of each predictive task alongside numerical results, including the entries in the contingency matrix for both the SVM and persistence predictive approaches. Figure 11 accompanies Table 5, graphically showing the TSS results (*y* axis) for each predictive task (*x* axis). To improve the visual representation, data point sizes, locations on the *y* axis, and color each reflect the magnitude of the TSS.

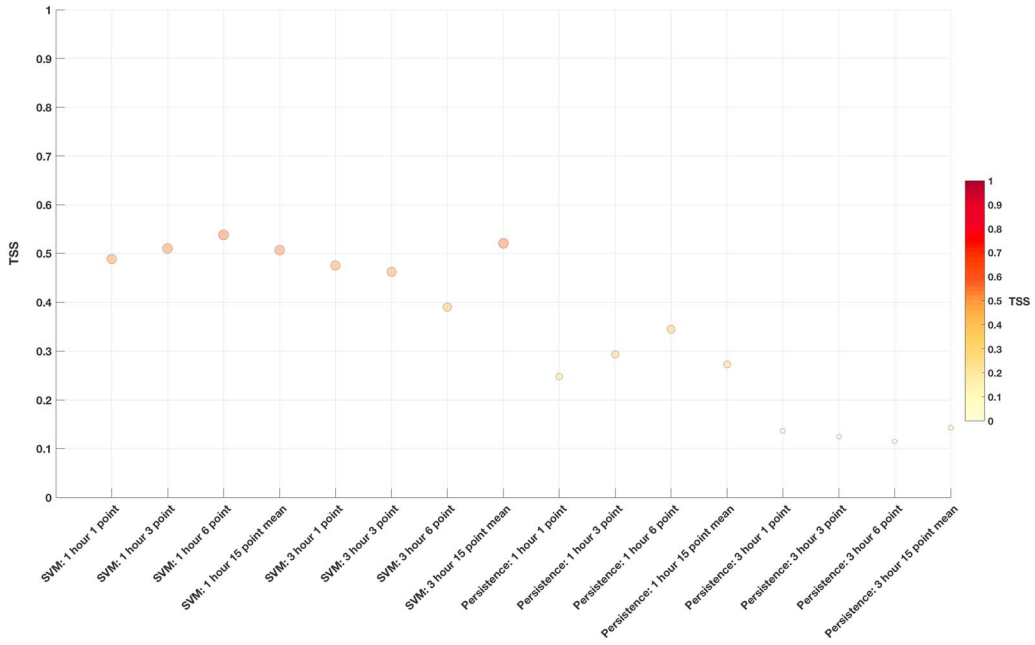

Figure 11. TSS results for the SVM model and persistence predictions evaluated for different predictive tasks over the period of 15–25 January 2016. To improve the visual representation, data point sizes, locations on the y-axis, and color each reflect the magnitude of the TSS. Refer to Table 5 for numerical results and definitions of the predictive tasks. SVM = support vector machine; TSS = true skill score.

From Figure 11 we find that the persistence predictive capability falls off drastically between 1- and 3-hr prediction times (only slightly exceeding random prediction skill for the 3-hr task), while the SVM model only experiences a minor decrease in skill. This trend is expected to continue with increasing prediction time, suggesting that the SVM approach offers the potential for skillful prediction at extended prediction times whereas persistence rapidly becomes ineffective. In general across all predictive tasks the SVM models outperform the persistence models, in most cases reaching values 2 times greater than persistence for the same predictive task (e.g., the TSS for the SVM model during this period for the *1-hr*, *1-point* task is 0.5 and that of persistence is 0.25). Finally, the SVM models for the *15-point-mean* predictive tasks show high skill. This may be a more appropriate predictive task for many users of GNSS signals given that the prediction is based on average scintillation conditions over a time period (15 min) significant to the operation of systems that require GNSS availability.

Ultimately, the most appropriate predictive task is application dependent, and end user needs should dictate predictive task selection.

6. Discussion

We have demonstrated the potential for skillful high-latitude ionospheric phase scintillation using an SVM model and benchmarked the performance using the TSS, which can be used across different predictive models to compare performance largely independent of the details of the input data (Bloomfield et al., 2012). We have not identified an optimal SVM model. Thus, it is important to consider how this model could be improved. First, knowledge of the space weather system can be utilized. Figure 10b showed that the SVM model with large class weight ratios is prone to overpredict scintillation occurrence (reference the large number of FPs for a model trained with a class weight ratio of 1:50). We also found the MLAT-MLT locations of the FPs peaked in the dayside cusp and premidnight auroral regions. One way to mitigate the shortcoming is to incorporate additional information pertaining to cusp and substorm phenomena into the input feature space. For instance, several studies (e.g., Jin et al., 2016; Watson et al., 2016a) found that scintillation in the cusp region may be closely related to the convection of ionospheric irregularities and would, therefore, benefit from convection data that could be provided by the Super Dual Auroral Radar Network (SuperDARN; Ruohoniemi & Baker, 1998) or an ionospheric convection model such as the SuperDARN assimilative model (Cousins et al., 2013). Field-aligned currents have been shown to relate closely to substorm behavior (McGranaghan,

Mannucci, et al., 2017; Murphy et al., 2013, and references therein), and field-aligned currents data from the Active Magnetosphere and Planetary Electrodynamics Response Experiment (AMPERE; Anderson et al., 2014) may, therefore, improve performance for this important scintillation-causing phenomenon. Further, perturbations in the Earth's magnetic field caused by horizontal electrical currents in the ionosphere, which can be sensed by ground-based magnetometers, provide a proxy for the dynamics of the magnetosphere and magnetosphere-ionosphere coupling (Kamide et al., 1981). The horizontal currents may be closely related to high-latitude phase scintillation (Prikryl et al., 2017). Therefore, the Super Magnetometer Initiative (SuperMAG; Gjerloev & Hoffman, 2012), coordinating data from more than 100 ground-based magnetometers, may be an important source of data as well. The SuperDARN, AMPERE, and SuperMAG data will also likely be beneficial to overall model improvement. Exploration of additional input features should be addressed through the lens of identifying the data that provide the greatest potential for discriminating among the scintillation and no scintillation classes, particularly for the cases that are most difficult to classify.

Next, sophisticated approaches to improve machine learning model performance from the machine learning research community can also contribute to drastic improvements. For instance, we have used random data selection to produce the data samples with which to train our SVM models. However, *smart* subsampling methods, such as the Synthetic Minority Oversampling Technique (Chawla et al., 2002), may prove more effective in selecting data to train more capable models.

Despite the fact that the SVM models that we presented are not in a strict sense *optimal*, they do exhibit skill and additionally provide new understanding for high-latitude phase scintillation. Referring to Figures 4 and 5 we found that several input features contributed little to the overall predictive capability of the SVM model. For instance, individual solar wind parameters (e.g., IMF B_y) had relatively low univariate scores and did little to increase the predictive capability of the SVM model. This, coupled with the large univariate scores of geomagnetic activity indices (e.g., *Kp*) and coupling functions, suggests that not only are combinations of data more effective to improve predictive skill for high-latitude ionospheric scintillation but also that single variables may not sufficiently distinguish ionospheric phenomenon. The latter points to the need to evolve approaches that traditionally attempt to understand *characteristics* (i.e., repeatable behavior) of ionospheric phenomena by examining ionospheric behavior as a function of a single variable (such as IMF clock angle). This is consistent with the conclusions drawn by Gjerloev et al. (2018) that it is unlikely that the ionosphere can be described by cause and effect relationships with individual solar wind variables nor that small subsets of parameters can accurately encode the necessary complexity. Our approach using SVMs allows an exploration of nonlinear combinations of a large number of relevant parameters for the prediction of ionospheric scintillation. We suggest that this supports the conclusion that machine learning approaches can complement traditional approaches to improve understanding and prediction of the geospace environment (McGranaghan, Bhatt, et al., 2017).

Our results also suggest new information regarding ionospheric predictability. Figure 11 showed that persistence prediction is almost completely ineffective for prediction times longer than 1 hr for high-latitude phase scintillation (i.e., performs just better than random chance for the 3-hr predictive task). This suggests that the *memory* of the ionosphere in terms of high-latitude plasma irregularities is on the order of or shorter than hours, an important result given the outstanding question of the predictability of the ionosphere (Mannucci et al., 2016). In general, new scientific understanding has been created from our machine learning approach and is indicative of the capability for *explainable machine learning* through the fusion of data-driven approaches with scientific knowledge (Karpatne et al., 2017).

We acknowledge that machine learning is only one approach to high-latitude ionospheric scintillation prediction. We do not claim that this is the best or most appropriate approach, only that traditional approaches and models should be compared with and complemented by machine learning approaches, given the availability of vastly increased data volumes and computational power that support successful machine learning techniques. This work is intended to provide the foundation for comparing various approaches and integrating traditional and novel approaches to produce more accurate and capable prediction models. There are many machine learning approaches that are not explored in this work. We present one promising approach and position this model as a benchmark to support future exploration.

Finally, we chose the prediction of the phase SI, σ_Φ, as our predictive task due to the *actionable* information for GNSS users provided by such a prediction (e.g., Albanese et al., 2017; Prikryl et al., 2015; Sreeja et al., 2012; Strangeways, 2009), the availability of large volumes of SI data, and the body of literature surrounding this

task (e.g., Spogli et al., 2009; Prikryl et al., 2012; Rezende et al., 2009, and references therein). Other useful approaches to ionospheric scintillation prediction also exist. Conker et al. (2003), Aquino et al. (2009), and Sreeja (2016) point out that the variance on the GNSS receiver phase-locked loop is a critical parameter to the calculation of the navigation solution from GNSS signals. Therefore, another useful predictive task would be to predict the variance of the error at the output of the phase-locked loop at a given location (i.e., a *tracking jitter map* [Sreeja et al., 2011]). The approach outlined in this work could indeed be applied to predict tracking jitter maps, and is, therefore, adaptable to serve GNSS user needs.

6.1. Future Work

This study examined a focused, well-defined, albeit minimal use case to explore promising, novel data-driven approaches to high-latitude scintillation. We note that data-driven approaches inherently open massive exploration spaces. The high-latitude scintillation data that are provided by the CHAIN GNSS receivers and that have been curated into an open and usable database through this research provide an opportunity for widespread exploration of this space. We detail a few important avenues of that exploration here and note that follow-on work in these regards is ongoing.

Future work will address four areas of investigation:

1. additional input features (e.g., SuperDARN convection) and determination of the importance of these additional data to scintillation prediction;
2. additional methods of featurization (e.g., discriminant analysis);
3. additional methods to subsample data (more sophisticated than a simple random subsampling);
4. a broad spectrum of machine learning algorithms (e.g., neural networks).

Given that the SVM model produces a high number of false alarms, these investigations will be geared toward reducing the false alarm rate without sacrificing the high number of true scintillation predictions. This effort may benefit from a deeper investigation into evaluation metrics used in cooperation with the TSS.

7. Conclusions

We have addressed a critical and unresolved task in space weather: the prediction of high-latitude ionospheric phase scintillation. We used a data-driven approach, with a large volume of data from GNSS signals collected by the CHAIN ground receivers, to develop a novel machine learning method to predict the occurrence of phase scintillation at magnetic latitudes poleward of 45° . We chose the SVM algorithm and evaluated the predictive performance using this approach based on a robust metric, the TSS, which can be reliably used to compare between predictive methods, to establish our results as a benchmark for future efforts. To our knowledge, this is the first time a SVM model has been applied to predict high-latitude phase scintillation. We find that the SVM model consistently and significantly outperforms a persistence prediction, yielding TSS values double that of persistence for a 1-hr predictive task. The improvement is even greater for longer prediction times, for which persistence experiences significant degradation in predictive skill but the SVM exhibits only slight decreases between 1- and 3-hr predictive tasks. We use knowledge of the space weather system to investigate the capabilities and shortcomings of the SVM model and find that overprediction (i.e., predicting scintillation to occur more often than it actually does) is a challenge but may be able to be addressed through further optimization of the model. Additionally, improving prediction capability in critical areas such as the ionospheric cusp and premidnight auroral region (i.e., the statistical location of prominent substorm behavior) may require additional data to inform the model such as the SuperDARN, the AMPERE, and the SuperMAG. A useful aspect of this work is that it is readily extensible to metrics other than the TSS, which may benefit application-specific investigations.

Acknowledgments
This research was supported by the NASA Living With a Star Jack Eddy Postdoctoral Fellowship Program, administered by the University Corporation for Atmospheric Research and coordinated through the Cooperative Programs for the Advancement of Earth System Science (CPAESS). R. M. M. was also partially supported by the JPL Data Science Working Group. Portions of this research were carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration and funded by the Data Science Working Group Pilot Project *Stretching Global Navigation Satellite Systems (GNSS) signals for Space Weather discovery*. The authors gratefully acknowledge the input of P.T. Jayachandran, PI of the CHAIN network for guidance on data interpretation, details of phase scintillation/variation at high latitudes, and comments on the manuscript. CHAIN is supported by the Canadian Foundation for Innovation and the New Brunswick Innovation Foundation. CHAIN operation is conducted in collaboration with the Canadian Space Agency. Science funding is provided by the Natural Sciences and Engineering Research Council of Canada.
Data used in this work are available from NASA's Coordinated Data Analysis Web (CDAWeb, https://cdaweb.sci.gsfc.nasa.gov/), the Oval Variation, Assessment, Tracking, Intensity, and Online Nowcasting Prime (OVATION Prime, http://sourceforge.net/projects/ovation-prime/), and the CHAIN (http://chain.physics.unb.ca/chain/). We openly and freely provide sample software to produce the results shown in this manuscript through a FigShare Project with the same name as this paper and available at https://doi.org/10.6084/m9.figshare.6813143 (McGranaghan et al., 2018a). Data from this work are also provided through the FigShare Project at https://doi.org/10.6084/m9.figshare.6813131 (McGranaghan et al., 2018b).

References

Aarons, J., & Basu, S. (1994). Ionospheric amplitude and phase fluctuations at the GPS frequencies. In *Proceedings of the 7th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1994)*, The Institute of Navigation, pp. 1569–1578.

Albanese, C., Rodriguez, F., Ronchini, R., Di Rollo, S., Berrilli, F., Cristaldi, A., et al. (2017). The ionosphere prediction service. *Proceedings of the International Astronomical Union*, *13*(S335), 352–354. https://doi.org/10.1017/S174392131800025X

Alfonsi, L., Spogli, L., De Franceschi, G., Romano, V., Aquino, M., Dodson, A., & Mitchell, C. N. (2011). Bipolar climatology of GPS ionospheric scintillation at solar minimum. *Radio Science*, *46*, RS0D05. https://doi.org/10.1029/2010RS004571

Anderson, D. (1973). A theoretical study of the ionospheric F region equatorial anomaly II. Results in the American and Asian sectors. *Planetary and Space Science*, *21*(3), 421–442. https://doi.org/10.1016/0032-0633(73)90041-X

Anderson, B. J., Korth, H., Waters, C. L., Green, D. L., Merkin, V. G., Barnes, R. J., & Dyrud, L. P. (2014). Development of large-scale Birkeland currents determined from the Active Magnetosphere and Planetary Electrodynamics Response Experiment. *Geophysical Research Letters*, *41*, 3017–3025. https://doi.org/10.1002/2014GL059941

Aquino, M., Monico, J. F. G., Dodson, A. H., Marques, H., De Franceschi, G., Alfonsi, L., et al. (2009). Improving the GNSS positioning stochastic model in the presence of ionospheric scintillation. *Journal of Geodesy*, *83*(10), 953–966. https://doi.org/10.1007/s00190-009-0313-6

Barnes, G., Leka, K. D., Schrijver, C. J., Colak, T., Qahwaji, R., Ashamari, O. W., et al. (2016). A comparison of flare forecasting methods. I. Results from the 'All-Clear' workshop. *The Astrophysical Journal*, *829*(2), 89.

Basu, S., MacKenzie, E., & Basu, S. (1988). Ionospheric constraints on VHF/UHF communications links during solar maximum and minimum periods. *Radio Science*, *23*(3), 363–378. https://doi.org/10.1029/RS023i003p00363

Beutler, G., Rothacher, M., Schaer, S., Springer, T., Kouba, J., & Neilan, R. (1999). The International GPS Service (IGS): An interdisciplinary service in support of Earth sciences. *Advances in Space Research*, *23*(4), 631–653. https://doi.org/10.1016/S0273-1177(99)00160-X, satellite Dynamics, Orbit Analysis and Combination of Space Techniques.

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. (2012). Toward reliable benchmarking of solar flare forecasting methods. *The Astrophysical Journal Letters*, *747*(2), L41.

Bobra, M. G., & Couvidat, S. (2015). Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal*, *798*(2), 135.

Borovsky, J. E. (2013). Physical improvements to the solar wind reconnection control function for the Earth's magnetosphere. *Journal of Geophysical Research: Space Physics*, *118*, 2113–2121. https://doi.org/10.1002/jgra.50110

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 593463. https://doi.org/10.1023/A:1009715923555

Carrano, C. S., & Rino, C. L. (2016). A theory of scintillation for two-component power law irregularity spectra: Overview and numerical results. *Radio Science*, *51*, 789–813. https://doi.org/10.1002/2015RS005903

Chapelle, O. (2007). Training a support vector machine in the primal. *Neural Computation*, *19*(5), 1155–1178. https://doi.org/10.1162/neco.2007.19.5.1155

Chartier, A., Forte, B., Deshpande, K., Bust, G., & Mitchell, C. (2016). Three-dimensional modeling of high-latitude scintillation observations. *Radio Science*, *51*, 1022–1029. https://doi.org/10.1002/2015RS005889

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357.

Cherniak, I., Krankowski, A., & Zakharenkova, I. (2014). Observation of the ionospheric irregularities over the Northern Hemisphere: Methodology and service. *Radio Science*, *49*, 653–662. https://doi.org/10.1002/2014RS005433

Codrescu, M. V., Negrea, C., Fedrizzi, M., Fuller-Rowell, T. J., Dobin, A., Jakowsky, N., et al. (2012). A real-time run of the Coupled Thermosphere Ionosphere Plasmasphere Electrodynamics (CTIPe) model. *Space Weather*, *10*, S02001. https://doi.org/10.1029/2011SW000736

Conker, R. S., El-Arini, M. B., Hegarty, C. J., & Hsiao, T. (2003). Modeling the effects of ionospheric scintillation on GPS/Satellite-Based Augmentation System availability. *Radio Science*, *38*(1), 1001. https://doi.org/10.1029/2000RS002604

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Coster, A., & Komjathy, A. (2008). Space weather and the global positioning system. *Space Weather*, *6*, S06D04. https://doi.org/10.1029/2008SW000400

Cousins, E. D. P., Matsuo, T., & Richmond, A. D. (2013). SuperDARN assimilative mapping. *Journal of Geophysical Research: Space Physics*, *118*, 7954–7962. https://doi.org/10.1002/2013JA019321

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, *EC-14*(3), 326–334. https://doi.org/10.1109/PGEC.1965.264137

Cowley, S. W. H. (2013). *Magnetosphere-ionosphere interactions: A tutorial review*. Washington, DC: American Geophysical Union (AGU). pp. 91–106, https://doi.org/10.1029/GM118p0091

Deshpande, K. B., Bust, G. S., Clauer, C. R., Scales, W. A., Frissell, N. A., Ruohoniemi, J. M., et al. (2016). Satellite-beacon Ionospheric-scintillation Global Model of the upper Atmosphere (SIGMA) II: Inverse modeling with high-latitude observations to deduce irregularity physics. *Journal of Geophysical Research: Space Physics*, *121*, 9188–9203. https://doi.org/10.1002/2016JA022943

Dierendonck, A. J. V., & Arbesser-Rastburg, B. (2001). Measuring Ionospheric scintillation in the equatorial region over Africa, including measurements from SBAS geostationary satellite signals. In *Proceedings of the 17th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2004)*, pp. 316–324.

Dierendonck, A. V., Klobuchar, J., & Hua, Q. (1997). Ionospheric scintillation monitoring using commercial single frequency C/A code receivers. In *Proceedings of the 6th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1993)*, ION GPS, pp. 1333–1342.

Duderstadt, J. J. (2001). *Issues for science and engineering researchers in the digital age*. Washington, DC: National Academy Press.

Forte, B. (2012). Analysis of the PLL phase error in presence of simulated ionospheric scintillation events. *Radio Science*, *47*, RS3006. https://doi.org/10.1029/2011RS004790

Forte, B., Coleman, C., Skone, S., Häggström, I., Mitchell, C., Da Dalt, F., et al. (2016). Identification of scintillation signatures on GPS signals originating from plasma structures detected with EISCAT incoherent scatter radar along the same line of sight. *Journal of Geophysical Research: Space Physics*, *122*, 916–931. https://doi.org/10.1002/2016JA023271

Gjerloev, J. W., & Hoffman, R. A. (2012). The large-scale current system during auroral substorms. *Journal of Geophysical Research*, *119*, 4591–4606. https://doi.org/10.1002/2013JA019176

Gjerloev, J. W., Waters, C. L., & Barnes, R. J. (2018). Deriving global convection maps from SuperDARN measurements. *Journal of Geophysical Research: Space Physics*, *123*, 2902–2915. https://doi.org/10.1002/2017JA024543

Gu, Q., Li, Z., & Han, J. (2011). Generalized Fisher score for feature selection. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI 11, AUAI Press, Arlington, Virginia, United States, pp. 266–273.

Habarulema, J. B., McKinnell, L.-Å., & Opperman, B. D. L. (2011). Regional GPS TEC modeling: Attempted spatial and temporal extrapolation of TEC using neural networks. *Journal of Geophysical Research*, *116*, A04314. https://doi.org/10.1029/2010JA016269

Hanson, W. B., & Moffett, R. J. (1966). Ionization transport effects in the equatorial F region. *Journal of Geophysical Research*, *71*(23), 5559–5572. https://doi.org/10.1029/JZ071i023p05559

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York, NY, USA: Springer Series in Statistics, Springer New York Inc.

Jayachandran, P. T., Langley, R. B., MacDougall, J. W., Mushini, S. C., Pokhotelov, D., Hamza, A. M., et al. (2009). Canadian High Arctic Ionospheric Network (CHAIN). *Radio Science*, *44*, RS0A03. https://doi.org/10.1029/2008RS004046

Jiao, Y., Hall, J. J., & Morton, Y. T. (2017). Automatic equatorial GPS amplitude scintillation detection using a machine learning algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, *53*(1), 405–418. https://doi.org/10.1109/TAES.2017.2650758

Jiao, Y., Morton, Y. T., Taylor, S., & Pelgrum, W. (2013). Characterization of high-latitude ionospheric scintillation of GPS signals. *Radio Science*, *48*, 698–708. https://doi.org/10.1002/2013RS005259

Jin, Y., Moen, J. I., & Miloch, W. J. (2016). On the collocation of the cusp aurora and the GPS phase scintillation: A statistical study. *Journal of Geophysical Research: Space Physics*, *120*, 9176–9191. https://doi.org/10.1002/2015JA021449

Jonas, E., Bobra, M. G., Shankar, V., Hoeksema, J. T., & Recht, B. (2017). Flare prediction using photospheric and coronal image data. ArXiv e-prints.

Kamide, Y., Richmond, A. D., & Matushita, S. (1981). Estimation of ionospheric electric fields, ionospheric currents, and field-aligned currents from ground magnetic records. *Journal of Geophysical Research*, *86*(A2), 801–813. https://doi.org/10.1029/JA086iA02p00801

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pp. 372–378. https://doi.org/10.1109/SAI.2014.6918213

Kintner, P. M. (2001). Fading timescales associated with GPS signals and potential consequences. *Radio Science*, *36*(4), 731–743. https://doi.org/10.1029/1999RS002310

Kintner, P. M., Ledvina, B. M., & de Paula, E. R. (2007). GPS and ionospheric scintillations. *Space Weather*, *5*, S09003. https://doi.org/10.1029/2006SW000260

Komjathy, A. (1997). Global ionospheric total electron content mapping using the global positioning system (Ph.D. thesis), THE UNIVERSITY OF NEW BRUNSWICK (CANADA, Fredericton, New Brunswick, Canada.

Leka, K. D., & Barnes, G. (2003). Photospheric magnetic field properties of flaring versus flare-quiet active regions II. Discriminant analysis. *The Astrophysical Journal*, *595*, 1296–1306. https://doi.org/10.1086/377512

Lima, G. R. T., Stephany, S., Paula, E. R., Batista, I. S., & Abdu, M. A. (2015). Prediction of the level of ionospheric scintillation at equatorial latitudes in Brazil using a neural network. *Space Weather*, *13*, 446–457. https://doi.org/10.1002/2015SW001182

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. CoRR, abs/1305.1707.

Lu, G., Holzer, T. E., Lummerzheim, D., Ruohoniemi, J. M., Stauning, P., Troshichev, O., et al. (2002). Ionospheric response to the interplanetary magnetic field southward turning: Fast onset and slow reconfiguration. *Journal of Geophysical Research*, *107*, 1153. https://doi.org/10.1029/2001JA000324

Machol, J. L., Green, J. C., Redmon, R. J., Viereck, R. A., & Newell, P. T. (2012). Evaluation of OVATION Prime as a forecast model for visible aurorae. *Space Weather*, *10*, S03005. https://doi.org/10.1029/2011SW000746

Mannucci, A. J., Hagan, M. E., Vourlidas, A., Huang, C. Y., Verkhoglyadova, O. P., & Deng, Y. (2016). Scientific challenges in thermosphere-ionosphere forecasting—Conclusions from the October 2014 NASA JPL community workshop. *Journal of Space Weather and Space Climate*, *6*, E01. https://doi.org/10.1051/swsc/2016030

Manzato, A. (2005). The use of sounding-derived indices for a neural network short-term thunderstorm forecast. *Weather and Forecasting*, *20*(6), 896–917. https://doi.org/10.1175/WAF898.1

McGranaghan, R. M., Bhatt, A., Matsuo, T., Mannucci, A. J., Semeter, J. L., & Datta-Barua, S. (2017). Ushering in a new frontier in geospace through data science. *Journal of Geophysical Research: Space Physics*, *122*, 12,586–12,590. https://doi.org/10.1002/2017JA024835

McGranaghan, R. M., Mannucci, A. J., & Forsyth, C. (2017). A comprehensive analysis of multiscale field-aligned currents: Characteristics, controlling parameters, and relationships. *Journal of Geophysical Research: Space Physics*, *122*, 11,931–11,960. https://doi.org/10.1002/2017JA024742

McGranaghan, R., Mannucci, A., Mattmann, C., Wilson, B., & Chadwick, R. (2018a). Jupyter notebook script to demonstrate the use of the machine learning databases and analysis for Journal of Geophysical Research: Space Physics manuscript: "New capabilities for prediction of high-latitude ionospheric scintillation: A novel approach with machine learning." (Version 1). figshare. https://doi.org/10.6084/m9.figshare.6813143.v1

McGranaghan, R., Mannucci, A., Mattmann, C., Wilson, B., & Chadwick, R. (2018b). Machine learning databases used for Journal of Geophysical Research: Space Physics manuscript: "New capabilities for prediction of high-latitude ionospheric scintillation: A novel approach with machine learning." (Version 1). figshare. https://doi.org/10.6084/m9.figshare.6813131.v1

McGranaghan, R. M., Mannucci, A. J., Verkhoglyadova, O., & Malik, N. (2017). Finding multiscale connectivity in our geospace observational system: Network analysis of total electron content. *Journal of Geophysical Research: Space Physics*, *122*, 7683–7697. https://doi.org/10.1002/2017JA024202

McPherron, R. L., & Siscoe, G. (2004). Probabilistic forecasting of geomagnetic indices using solar wind air mass analysis. *Space Weather*, *2*, S01001. https://doi.org/10.1029/2003SW000003

Mezaoui, H., Hamza, A. M., & Jayachandran, P. T. (2014). Investigating high-latitude ionospheric turbulence using Global Positioning System data. *Geophysical Research Letters*, *41*, 6570–6576. https://doi.org/10.1002/2014GL061331

Mitchell, C. N., Alfonsi, L., De Franceschi, G., Lester, M., Romano, V., & Wernik, A. W. (2005). GPS TEC and scintillation measurements from the polar ionosphere during the October 2003 storm. *Geophysical Research Letters*, *32*, L12S03. https://doi.org/10.1029/2004GL021644

Mourad, S., Tewfik, A., & Vikalo, H. (2017). Data subset selection for efficient SVM training. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 833–837. https://doi.org/10.23919/EUSIPCO.2017.8081324

Mrak, S., Semeter, J., Hirsch, M., Starr, G., Hampton, D., Varney, R. H., et al. (2017). Field-aligned GPS scintillation: Multisensor data fusion. *Journal of Geophysical Research: Space Physics*, *123*, 974–992. https://doi.org/10.1002/2017JA024557

Muella, M. T. A. H., Kherani, E. A., de Paula, E. R., Cerruti, A. P., Kintner, P. M., Kantor, I. J., et al. (2010). Scintillation-producing Fresnel-scale irregularities associated with the regions of steepest TEC gradients adjacent to the equatorial ionization anomaly. *Journal of Geophysical Research*, *115*, A03301. https://doi.org/10.1029/2009JA014788

Murphy, K. R., Mann, I. R., Rae, I. J., Waters, C. L., Frey, H. U., Kale, A., et al. (2013). The detailed spatial structure of field-aligned currents comprising the substorm current wedge. *Journal of Geophysical Research: Space Physics*, *118*, 7714–7727. https://doi.org/10.1002/2013JA018979

Newell, P. T., & Meng, C.-I. (1988). The cusp and the cleft/boundary layer: Low-altitude identification and statistical local time variation. *Journal of Geophysical Research*, *93*(A12), 14,549–14,556. https://doi.org/10.1029/JA093iA12p14549

Newell, P. T., Sotirelis, T., Liou, K., Meng, C.-I., & Rich, F. J. (2007). A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables. *Journal of Geophysical Research*, *112*, A01206. https://doi.org/10.1029/2006JA012015

Newell, P. T., Sotirelis, T., & Wing, S. (2009). Diffuse, monoenergetic, and broadband aurora: The global precipitation budget. *Journal of Geophysical Research*, *114*, A09207. https://doi.org/10.1029/2009JA014326

Newell, P. T., Sotirelis, T., & Wing, S. (2010). Seasonal variations in diffuse, monoenergetic, and broadband aurora. *Journal of Geophysical Research*, *115*, A03216. https://doi.org/10.1029/2009JA014805

Ng, A. (2018). Lecture notes in Stanford Coursera machine learning course: Machine learning. https://www.coursera.org/learn/machine-learning

Norman, R., Carter, B., Bennett, J., Le Marshall, J., Hearne, J., & Zhang, K. (2016). Australian space research program—Platform technologies for space, atmosphere and climate project: Selected innovations. In R. S. Anderssen, et al. (Eds.), *Applications + practical conceptualization + mathematics = fruitful innovation* pp. 159–174). Tokyo: Springer Japan.

Oksavik, K., Søraas, F., Moen, J., & Burke, W. J. (2000). Optical and particle signatures of magnetospheric boundary layers near magnetic noon: Satellite and ground-based observations. *Journal of Geophysical Research*, *105*(A12), 27,555–27,568. https://doi.org/10.1029/1999JA000237

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2078195.

Prikryl, P., Ghoddousi-Fard, R., Viljanen, A., Weygand, J. M., Kunduri, B. S. R., Thomas, E. G., et al. (2017). GPS phase scintillation and auroral electrojet currents during geomagnetic storms of March 17, 2013 and 2015. In *2017 XXXIInd General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS)*, pp. 1–4. https://doi.org/10.23919/URSIGASS.2017.8105125

Prikryl, P., Jayachandran, P. T., Chadwick, R., & Kelly, T. D. (2015). Climatology of GPS phase scintillation at northern high latitudes for the period from 2008 to 2013. *Annales Geophysicae*, *33*(5), 531–545. https://doi.org/10.5194/angeo-33-531-2015

Prikryl, P., Jayachandran, P. T., Mushini, S. C., & Richardson, I. G. (2012). Toward the probabilistic forecasting of high-latitude GPS phase scintillation. *Space Weather*, *10*, S08005. https://doi.org/10.1029/2012SW000800

Prikryl, P., Vadakke Veettil, S., Aquino, M., & Jayachandran, P. (2013). Probabilistic forecasting of ionospheric scintillation and GNSS receiver signal tracking performance at high latitudes. *Annals of Geophysics*, *56*(2).

Redmon, R. J., Anderson, D., Caton, R., & Bullett, T. (2010). A forecasting ionospheric real-time scintillation tool (FIRST). *Space Weather*, *8*, S12003. https://doi.org/10.1029/2010SW000582

Ren, X., Zhang, X., Xie, W., Zhang, K., Yuan, Y., & Li, X. (2016). Global ionospheric modelling using multi-GNSS: BeiDou, Galileo, GLONASS and GPS. *Scientific Reports*, *6*(1), 33499. https://doi.org/10.1038/srep33499

Rezende, L. F. C., de Paula, E. R., Stephany, S., Kantor, I. J., Muella, M. T. A. H., de Siqueira, P. M., & Correa, K. S. (2009). Survey and prediction of the ionospheric scintillation using data mining techniques. *Space Weather*, *8*, S06D09. https://doi.org/10.1029/2009SW000532

Ridley, A. J., Lu, G., Clauer, C. R., & Papitashvili, V. O. (1998). A statistical study of the ionospheric convection response to changing interplanetary magnetic field conditions using the assimilative mapping of ionospheric electrodynamics technique. *Journal of Geophysical Research*, *103*(A3), 4023–4039. https://doi.org/10.1029/97JA03328

Rizos, C., Montenbruck, O., Weber, G. R., Weber, R., & Neilan, U. H. (2013). The IGS MGEX experiment as a milestone for a comprehensive multi-GNSS service. In *Proceedings of the ION 2013 Pacific PNT Meeting*, The Institute of Navigation, pp. 289–295.

Ruohoniemi, J. M., & Baker, K. B. (1998). Large-scale imaging of high-latitude convection with Super Dual Auroral Radar Network HF radar observations. *Journal of Geophysical Research*, *103*(A9), 20,797–20,811. https://doi.org/10.1029/98JA01288

Sadlier, G., Flytkjær, R., Sabri, F., & Herr, D. (2017). *The economic impact on the UK of a sisruption to GNSS*. Innovate UK, UK Space Agency: Royal Institute of Navigation.

Schölkopf, B., Tsuda, K., & Vert, J.-P. (2004). *Kernel methods in computational biology*. Cambridge, Massachusetts: MIT Press.

Secan, J. (1995). An improved model of equatorial scintillation. *Radio Science*, *30*(3), 607–617. https://doi.org/10.1029/94RS03172

Secan, J. A., Bussey, R. M., Fremouw, E. J., & Basu, S. (1997). High-latitude upgrade to the wideband ionospheric scintillation model. *Radio Science*, *32*(4), 1567–1574. https://doi.org/10.1029/97RS00453

Semeter, J., Hirsch, M., Lind, F., Coster, A., Erickson, P., & Pankratius, V. (2016). GNSS-ISR data fusion: General framework with application to the high-latitude ionosphere. *Radio Science*, *51*, 118–129. https://doi.org/10.1002/2015RS005794

Semeter, J., Mrak, S., Hirsch, M., Swoboda, J., Akbari, H., Starr, G., et al. (2017). GPS signal corruption by the discrete aurora: Precise measurements from the Mahali experiment. *Geophysical Research Letters*, *44*, 9539–9546. https://doi.org/10.1002/2017GL073570

Septentrio (2015). *PolaRxS application manual*. Leuven, Belgium: Septentrio Satellite Navigation.

Shepherd, S. G. (2014). Altitude-adjusted corrected geomagnetic coordinates: Definition and functional approximations. *Journal of Geophysical Research: Space Physics*, *119*, 7501–7521. https://doi.org/10.1002/2014JA020264

Siscoe, G., & Solomon, S. C. (2006). Aspects of data assimilation peculiar to space weather forecasting. *Space Weather*, *4*, S04002. https://doi.org/10.1029/2005SW000205

Spogli, L., Alfonsi, L., De Franceschi, G., Romano, V., Aquino, M. H. O., & Dodson, A. (2009). Climatology of GPS ionospheric scintillations over high and mid-latitude European regions. *Annales Geophysicae*, *27*(9), 3429–3437. https://doi.org/10.5194/angeo-27-3429-2009

Sreeja, V. (2016). Impact and mitigation of space weather effects on GNSS receiver performance. *Geoscience Letters*, *3*(1), 24. https://doi.org/10.1186/s40562-016-0057-0

Sreeja, V., Aquino, M., Elmas, Z. G., & Forte, B. (2012). Correlation analysis between ionospheric scintillation levels and receiver tracking performance. *Space Weather*, *10*, S06005. https://doi.org/10.1029/2012SW000769

Sreeja, V. V., Aquino, M., Forte, B., Elmas, Z., Hancock, C., De Franceschi, G., et al. (2011). Tackling ionospheric scintillation threat to GNSS in Latin America. *Journal of Space Weather and Space Climate*, *1*(1), A05. https://doi.org/10.1051/swsc/2011005

Steenburgh, R. A., Biesecker, D. A., & Millward, G. H. (2014). From predicting solar activity to forecasting space weather: Practical examples of research-to-operations and operations-to-research. *Solar Physics*, *289*(2), 675–690. https://doi.org/10.1007/s11207-013-0308-6

Strangeways, H. J. (2009). Determining scintillation effects on GPS receivers. *Radio Science*, *44*, RS0A36. https://doi.org/10.1029/2008RS004076

Su, Y., Datta-Barua, S., Bust, G. S., & Deshpande, K. B. (2017). Distributed sensing of ionospheric irregularities with a GNSS receiver array. *Radio Science*, *52*, 988–1003. https://doi.org/10.1002/2017RS006331

Tzelepis, C., & Carreno, A. (2016). Concepts for NASA's Communication and Navigation Architecture in Near Earth and Deep Space Domains; Strategies for affordable and scalable implementation of next generation relay systems with improved mission experience and reduced burden. In *34th AIAA International Communications Satellite Systems Conference*. https://doi.org/10.2514/6.2016-5706

Uwamahoro, J. C., & Habarulema, J. B. (2015). Modelling total electron content during geomagnetic storm conditions using empirical orthogonal functions and neural networks. *Journal of Geophysical Research: Space Physics*, *120*, 11,000–11,012. https://doi.org/10.1002/2015JA021961

Wang, J., Morton, Y. J., & Hampton, D. (2017). New results on ionospheric irregularity drift velocity estimation using multi-GNSS space-receiver array during high-latitude phase scintillation. *Radio Science*, *53*, 228–240. https://doi.org/10.1002/2017RS006470

Wang, Y., Zhang, Q.-H., Jayachandran, P. T., Moen, J., Xing, Z.-Y., Chadwick, R., et al. (2018). Experimental evidence on the dependence of the standard GPS phase scintillation index on the ionospheric plasma drift around noon sector of the polar ionosphere. *Journal of Geophysical Research: Space Physics*, *123*, 2370–2378. https://doi.org/10.1002/2017JA024805

Watson, C., Jayachandran, P. T., & MacDougall, J. W. (2016a). Characteristics of GPS TEC variations in the polar cap ionosphere. *Journal of Geophysical Research: Space Physics*, *121*, 4748–4768. https://doi.org/10.1002/2015JA022275

Watson, C., Jayachandran, P. T., & MacDougall, J. W. (2016b). GPS TEC variations in the polar cap ionosphere: Solar wind and IMF dependence. *Journal of Geophysical Research: Space Physics*, *121*, 9030–9050. https://doi.org/10.1002/2016JA022937

Wernik, A. W. (1997). Wavelet transform of nonstationary ionospheric scintillation. *Acta Gephysica Polonica*, *45*, 237–253.

Wernik, A., Secan, J., & Fremouw, E. (2003). Ionospheric irregularities and scintillation. *Advances in Space Research*, *31*(4), 971–981. https://doi.org/10.1016/S0273-1177(02)00795-0

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques, second edition (Morgan Kaufmann series in data management systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Yeh, K. C., & Liu, C.-H. (1982). Radio wave scintillations in the ionosphere. *Proceedings of the IEEE*, *70*(4), 324–360. https://doi.org/10.1109/PROC.1982.12313

Zhang, X., Wu, Y., Wang, L., & Li, R. (2016). Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(1), 53–76. https://doi.org/10.1111/rssb.12100

Zou, Y., Nishimura, Y., Lyons, L. R., Donovan, E. F., Shiokawa, K., Ruohoniemi, J. M., et al. (2015). Polar cap precursor of nightside auroral oval intensifications using polar cap arcs. *Journal of Geophysical Research: Space Physics*, *120*, 10,698–10,711. https://doi.org/10.1002/2015JA021816