Space Weather

RESEARCH ARTICLE
10.1029/2020SW002555

On the Assessment of Daily Equatorial Plasma Bubble Occurrence Modeling and Forecasting

B. A. Carter1, J. L. Currie1, T. Dao1, E. Yizengaw2, J. Retterer3, M. Terkildsen4, K. Groves3, and R. Caton5

1SPACE Research Centre, School of Science, RMIT University, Melbourne, Victoria, Australia, 2Space Science Application Laboratory, The Aerospace Corporation, El Segundo, CA, USA, 3Institute for Scientific Research, Boston College, Boston, MA, USA, 4Space Weather Services, Bureau of Meteorology, Sydney, New South Wales, Australia, 5Space Vehicles Directorate, Air Force Research Laboratory, Albuquerque, NM, USA

Abstract
Predicting the daily variability of Equatorial Plasma Bubbles (EPBs) is an ongoing scientific challenge. Various methods for predicting EPBs have been developed, however, the research community is yet to scrutinize the methods for evaluating and comparing these prediction models/techniques. In this study, 12 months of co-located GPS and UHF scintillation observations spanning South America, Atlantic/Western Africa, Southeast Asia, and Pacific sectors are examined. Various assessment metrics are explored, including the use of significance testing on skill scores for threshold selection. The sensitivity of these skill scores to data set type (i.e., GPS versus UHF) and data set size (30, 50, 60, and 90 days/events) is also investigated. It is shown that between 50 and 90 days is required to achieve a statistically significant skill score. Methods for conducting model-model comparisons are also explored, including the use of model “sufficiency.” However, it is shown that the results of model-model comparisons must be carefully interpreted and can be heavily dependent on the data set used. It is also demonstrated that the observation data set must exhibit an appropriate level of daily EPB variability in order to assess the true strength of a given model/technique. Other limitations and considerations on assessment metrics and future challenges for EPB prediction studies are also discussed.

1. Introduction
Convective Ionospheric Storms, Equatorial Plasma Bubbles (EPBs), Equatorial Spread F (ESF), and equatorial ionospheric scintillation events are all associated with the same ionospheric plasma disturbance that naturally occurs in the nighttime low-latitude ionosphere (Kelley et al., 2011; Kintner et al., 2007). After sunset, the Generalized Rayleigh-Taylor (R-T) plasma instability, accompanied by existing small-scale plasma perturbations, facilitates the incursion of low-density plasma from the bottom-side F layer into the high-density plasma at higher altitudes, commonly referred to as EPBs. The plasma density gradients generated as a consequence of this larger perturbation cause energy to cascade to smaller scales, generating plasma irregularities with scale-lengths down to meters (e.g., Kelley et al., 1981). These smaller-scale magnetic field-aligned irregularities cause radar/ionosonde returns, which have the appearance of range and/or frequency spread on ionograms (e.g., Kelley et al., 1981). The resulting spatial plasma density gradients tend to focus and defocus coherent radio signals that traverse the region. When this occurs, random fluctuations in signal amplitude and phase, called “scintillations,” are induced and can result in complete loss-of-lock of satellite signals, lasting tens of seconds in severe cases (Kintner et al., 2007).

ESF/EPB research has generally been motivated by the curious aspects of the equatorial ionosphere as a natural plasma physics laboratory. However, the now heavy reliance of global society on satellite communications and navigation signals, and the adverse impacts that EPBs have on such signals, has strongly motivated research aimed at the prediction of these space weather events in recent times (e.g., de La Beaujardière, 2004; Kelley et al., 2011; Redmon et al., 2010; Retterer, 2005). Ionospheric scintillations are well-known to regularly impact terrestrial-based Global Navigation Satellite System (GNSS) applications; for example, agriculture, mining, defence, aviation, and construction. Ionospheric scintillation also impacts space-based GNSS applications; for example, satellite/shuttle tracking (e.g., Buchert et al., 2015; Goodman & Kramer, 2001; Xiong et al., 2016) and GNSS radio occultation (e.g., Brahmanandam et al., 2012; Carter et al., 2013;
Dymond, 2012; Straus et al., 2003) that is used in tropospheric weather, climate, and ionospheric monitoring (e.g., Anthes, 2011; Anthes et al., 2008). In fact, NASA’s Heliophysics Living with a Star Program highlighted “Physics-based scintillation forecasting capability” as “Strategic Science Area · 5” in their 10-year plan beyond 2015 (https://lwstrt.gsfc.nasa.gov/images/pdf/LWS_10YrVision_Oct2015_Final.pdf). At the time of writing, no space weather prediction agency in the world currently provides accessible ionospheric scintillation forecasts.

Research has been focused on improving our understanding of the physical processes that drive the formation of EPBs, as the first step in building a forecast capability. The seasonal/longitudinal variations in EPB occurrence has been a topic long under investigation (e.g., Aarons, 1993; Basu & Basu, 1985; Basu et al., 2002; Burke, Gentile, et al., 2004; Burke, Huang, et al., 2004; Carter, Retterer, Yizengaw, Groves, et al., 2004b; Dao et al., 2011; Nishioka et al., 2008; Retterer & Gentile, 2009; Su et al., 2008; Tsunoda, 1985; Yizengaw & Groves, 2018), and the consensus has formed that the longitudinal E-region conductivity gradient, which modulates the strength of the prereversal enhancement in the upward plasma drift (Abdu et al., 1981; Tsunoda, 1985), is the primary factor in EPB growth. EPB occurrence is understood to be high when this gradient is large (i.e., when the local magnetic field is closely aligned with the solar terminator) and low when this gradient is small (i.e., a large angle between the magnetic field and the terminator). With the exception of the African sector that does not appear to fully comply with this model (e.g., Yizengaw & Groves, 2018 and references therein), the seasonal/longitudinal variability (i.e., climatology) of EPB occurrence is largely understood.

The primary challenge in EPB prediction now lies in understanding and capturing day-to-day variability in the EPB occurrence. While it is widely understood that the R-T plasma instability is the primary mechanism causing EPBs (e.g., Haerendel et al., 1992; Sultan, 1996), the relative impacts of the physical parameters within that instability have been difficult to investigate, particularly in the context of background “seeds” that are also thought to play an important role in the plasma dynamics (e.g., Retterer & Roddy, 2014; Tsunoda, 2010a, 2010b; Yizengaw & Groves, 2018).

To that end, various studies have explored different ways to predict the onset of EPBs/ESF/scintillations as attempts to build a true forecasting capability. Anderson et al. (2004) utilized the understanding of a connection between the strength of the upward plasma drift after sunset and EPB growth to design and test a day-to-day forecasting technique that used ionosonde data in the South American sector. In that study, it was revealed that a threshold upward plasma drift of 20 m/s could be used to distinguish ESF days from non-ESF days, reporting a success rate of 84% versus 79% for the “persistence” forecast (i.e., what happened yesterday will happen today). This technique was further developed by Redmon et al. (2010) to utilize the h’F parameter in place of the upward plasma drift, which was shown to exhibit a clear threshold value at 1930 local time (LT) that varies linearly with solar activity. Anderson and Redmon (2017) continued work on this technique by extending its application to other longitude sectors and included an analysis of changes in EPB occurrence during the 2003 Halloween storms. It was found that the scintillation forecasting technique was successful in more than 80% of the cases investigated across South American, African, and Pacific longitude sectors.

While the upward plasma drift is a key parameter to consider, evidence suggests that “seed” perturbations could also be important (e.g., Huang, 2018). Thus, some research has focused on the use of seed perturbation detection for scintillation prediction (e.g., Sridharan et al., 2012, 2014, 2015; Yadav et al., 2017), while de Lima et al. (2015) developed a neural network that was trained and tested using ionosonde, scintillation, and solar flux data. Other researchers have adopted a more physics-based approach. For instance, Retterer et al. (2005) used vertical plasma drift observations to drive a regional physics-based model to capture changes in EPB growth for 7 days.

An alternative approach of using the Thermosphere Ionosphere Electrodynamics General Circulation Model (TIEGCM) to capture day-to-day variability in the R-T growth rate as a proxy for EPB occurrence has also been explored by Carter, Retterer, Yizengaw, Groves, et al. (2014), Carter, Retterer, Yizengaw, Wiens, et al. (2014), and Carter, Yizengaw, et al. (2014) using months of scintillation data from multiple longitude sectors. Wu (2015) investigated the seasonal/longitudinal variability in the TIEGCM R-T growth rate and found a good qualitative agreement with the EPB occurrence patterns from 15 years of in situ satellite observations previously reported by Gentile et al. (2006). This was followed by Wu (2017) with an
analysis of the impact of solar variability on the TIEGCM R-T growth rate. Shinagawa et al. (2018) analyzed the R-T growth rates calculated from outputs of the Ground-to-topside model of Atmosphere and Ionosphere for Aeronomy (GAIA) and compared them to 1 year of EPB occurrence data collected from Indonesia. In that analysis, variability was only enforced in the GAIA model from the lower atmosphere and not from geomagnetic activity. Nonetheless, the GAIA results exhibited a significant level of variability in the R-T growth rates, showing that variability from the lower atmosphere is an important factor.

Some other research has been focused on a few EPB case studies, which have been useful in revealing some underlying physics of EBP/ESF development (e.g., Carter et al., 2016; Hysell et al., 2019; Kelley & Retterer, 2008). While many of the studies above have considered data sets consisting of only a few days to years, some attention has been focused on the use of long-term data sets, for the purposes of studying long-term patterns and developing and testing climatological ESF/scintillation models (e.g., Aveiro & Hysell, 2010; Cervera et al., 2001; Forte & Radicella, 2005; Hysell & Burcham, 2002; Secan et al., 1995).

While all of the works on EPB prediction mentioned above have the same ultimate goal, the techniques, data, and even the assessment methods used in these studies have varied. Some studies have employed qualitative comparisons against observations, while others have used contingency tables/confusion matrices and forecast skill scores to quantify the success of their model/technique. The implication in the use of skill scores has been, of course, that a value reported by one study can be meaningfully compared against another (e.g., Carter, Yizengaw, et al., 2014). However, the meteorology literature has long warned that such comparisons do not necessarily reveal which model/technique is of “higher quality” (e.g., Ehrendorfer & Murphy, 1988; Murphy, 1993; Murphy & Ehrendorfer, 1987). Further, the data sets used in these studies were collected from one or multiple stations in various longitude sectors using different instruments and can range in coverage from only 1 day to several years. Finally, numerous studies have defined and used various parameter thresholds that were necessary in those works, including solar activity and scintillation levels, but detailed scrutiny of those threshold choices has not been performed. Given that EPB understanding and prediction is maturing, there is currently a need for the research community to attempt to standardize the evaluation and comparison of EPB prediction techniques/models, particularly if the universal goal of producing a reliable EPB forecasting capability is to be achieved. This sentiment is echoed within NASA’s Heliophysics Living with a Star Program report (https://lwstrt.gsfc.nasa.gov/images/pdf/LWS_10YrVision_Oct2015_Final.pdf), which specifies that “Assessment should go beyond case studies and model runs and should establish rigorous statistical quantification of limits of predictability.” Active discussions on these aspects are currently lacking in the literature.

In this study, a 12-month period of EPB occurrence is examined using a global network of co-located UHF-GPS receiver stations. The performance of the TIEGCM R-T growth rate in capturing the day-to-day EPB variability is assessed using several measures of forecast “skill,” and their limitations are explored. Finally, a systematic comparison is performed between the TIEGCM R-T growth rate and the persistence forecast for all stations, and the caveats in such comparisons are uncovered.

### 2. Data and Methodology

Recently, Halford et al. (2019) proposed the versatile Application Usability Level (AUL) framework that can be used to guide scientific research projects towards their intended applications. The AUL framework consists of nine different levels, ranging AUL 1 “Basic Research” to AUL 9 “Approved for on Demand Use.” Included in Halford et al. (2019) are several examples of research projects/applications across different areas of space physics, one of which is the project involving the translation of Carter, Retterer, Yizengaw, Groves, et al. (2014), Carter, Retterer, Yizengaw, Wiens, et al. (2014), and Carter, Yizengaw, et al. (2014) research into forecast operations at the Australian Bureau of Meteorology’s Space Weather Services. As detailed in that example (see section 4.4 in Halford et al., 2019), this project was rated as AUL 5 “Demonstration in relevant context,” citing the work by Carter, Retterer, Yizengaw, Wiens, et al. (2014) that achieved the AUL 5 milestones. The current study was motivated by the progression of the project towards meeting the milestones of AUL 6, which is the “Complete Validation” of this EPB occurrence forecast capability.

In this analysis, ionospheric scintillation data collected from 1 August 2013 until 25 July 2014 from a collection of GPS and UHF receiver stations in the Scintillation Network Decision Aid (SCINDA) network (Groves et al., 1997) were used. Figure 1a shows the locations of the SCINDA GPS (diamonds) and UHF (crosses).
stations, spanning South America, Western Africa/Atlantic, Southeast Asia, and the Pacific. The stations, from west to east, are Cuiaba (CBA-UHF data, East link) and Dourados (DRD-GPS data), which are considered as “co-located” in this analysis, Cape Verde (CVD, UHF, and GPS), Dakar (DKR, UHF, and GPS), Ascension Island (ASI, UHF, and GPS), Bangkok (BKK, UHF, and GPS), Taipei (TPU, UHF, and GPS), and Kwajalein Atoll (KWA, UHF, and GPS).

The GPS and UHF amplitude scintillation S4 index from equatorial/low-latitude stations is used to study the occurrence of postsunset EPBs. For each hour between 18 and 02 LT within all station data sets, the GPS S4 index data from satellite-to-ground links 30° above the horizon were used to calculate the 90th percentile, $S_{4p90}$. For the UHF S4 data, only one ground-to-satellite link with geostationary satellites was used to calculate $S_{4p90}$. As an indicator of the daily EPB/scintillation event variability, the daily maximum $S_{4p90}$ for both GPS and UHF was taken, $S_{4p90}\text{max}$. Upon close inspection, it was determined that the CBA UHF East link

![Figure 1](https://example.com/figure1.png)

**Figure 1.** (a) The locations of the SCINDA GPS (diamonds) and UHF (crosses) receiver stations used in this study. The solid line is the magnetic equator, and the dashed lines indicate magnetic latitudes $\pm 15^\circ$. (b) The daily maxima of the hourly 90th percentile of the GPS S4 index ($S_{4p90}$) versus date for all SCINDA stations in (a). The contours indicate the angle between the day-night terminator and the local magnetic field at sunset, and the gray cells indicate data gaps. (c) Same as (b) but for UHF $S_{4p90}$. (d) The daily maximum R-T growth rate versus date for all SCINDA stations in (a).
scintillation data collected from 26 February onwards was inconsistent with the GPS data and was deemed unreliable; these data were subsequently removed from further analysis.

This study also uses the outputs from the TIEGCM (Qian et al., 2014 and references therein). The TIEGCM is a self-consistent physics-based model that uses the Global Scale Wave Model (GSWM) (Hagan et al., 2001) to characterize tidal forcing on the lower altitudes, the F10.7 index to characterize solar activity input, and a choice of empirical high-latitude electric potential patterns (either “Heelis,” Heelis et al., 1982, or “Weimer,” Weimer, 2005, models) to characterize the magnetospheric input. In this study, both “Heelis” and “Weimer” high-latitude models were used, although the results displayed here have been obtained using the “Heelis” model, which takes the Kp index as an input.

The flux tube integrated R-T growth rate \( \gamma \) was derived by Sultan (1996) as

\[
\gamma = \frac{\Sigma_F}{\Sigma_P + \Sigma_F} \left( V_F - U_T^F - \frac{g_r}{v_{\text{eff}}} \right) K_F^F - R_T,
\]

where \( \Sigma_F \) is the Pedersen conductivity for E- and F-layer heights, \( V_F \) is the upward plasma drift, \( U_T^F \) is the Pedersen conductivity-weighted neutral wind perpendicular to the magnetic field in the meridional plane, \( g_r/v_{\text{eff}} \) is the altitude-corrected acceleration due to gravity divided by the effective ion-neutral collision frequency, \( K_F^F \) is the F-region flux tube electron content height gradient, and \( R_T \) is the recombination rate. Each of these terms are integrated along the magnetic flux tube (i.e., field line), so inter-hemispherical asymmetries are incorporated into the growth rate expression. Using the same methodology as Carter, Retterer, Yizengaw, Groves, et al. (2014), Carter, Retterer, Yizengaw, Wiens, et al. (2014), Carter, Yizengaw, et al. (2014), and Carter et al. (2016), the altitude profiles of the R-T growth rate were calculated for the magnetic meridians of each station in Figure 1a for every 15-min period between 18 and 02 LT from 1 August 2013 until 25 July 2014. In this methodology, the flux tubes from 90 to 500 km apex altitude were integrated along in 10-km steps using the International Reference Geomagnetic Field (Maus et al., 2005). For each integration location, the closest 8 model cells were averaged with a weighting of \( 1/r^2 \), where \( r \) is the linear distance to those cells (or their midpoints for some parameters). Also, the \( R_T \) term is assumed to be negligible. The result is an apex altitude profile of the flux tube integrated R-T growth rate along each station location meridian. To capture the daily variability in a similar manner to the S4 observations, the daily maximum R-T linear growth rate was used in the analysis.

To aid in the interpretation of the results that follow, it is necessary to make some brief remarks on the design of the TIEGCM and the associated implications for the resulting R-T growth rates in this study. Firstly, the coarse spatial grid employed by the TIEGCM prevents the inclusion of small-scale structures (i.e., “seeds”). Therefore, the following analysis assumes that seeds are invariant and do not influence the EPB occurrence patterns. While some previous modeling efforts have been successful in reproducing the observed EPB occurrence climatology using this assumption (e.g., Retterer & Gentile, 2009), seeding could be an important factor of daily variability that cannot be captured in this analysis. Further, the TIEGCM lower boundary is forced by a climatological wind-field model (GSWM) that does not include daily variability. Moreover, the model’s use of the Kp index limits its ability to capture small-scale temporal variations in the high-latitude input (e.g., during geomagnetic storms) that could be important for EPB development/suppression. In particular, the TIEGCM was not specifically designed to handle the penetration of high-latitude electric fields to low latitudes (e.g., Wang et al., 2008; Zaka et al., 2010). As such, magnetospheric penetration electric fields (e.g., Kikuchi et al., 2008) that influence the electrodynamics of EPB growth/suppression (e.g., Abdu, 2012; Huang, 2011) are not explicitly included in the modeling results (Carter et al., 2016, 2018). Therefore, the TIEGCM is considered to be a numerical simulation tool that can aid in understanding which physical phenomena are, and are not, dominant/important in controlling day-to-day EPB variability. The prime focus of this analysis is how to best quantify any model/technique’s “forecast” success, using the TIEGCM as an example.

3. Results

In this section, the GPS and UHF scintillation observations from all stations are reviewed and compared against the TIEGCM R-T growth rate results, followed by a detailed investigation into the use of skill
scores in quantifying modeling/forecast “success.” An analysis of the sensitivity of the skill scores to threshold selection and data set types and size is then conducted, followed by a comparison of the TIEGCM R-T growth rate against the persistence forecast using the concept of model “sufficiency.”

3.1. Qualitative Comparison of Scintillation Observations with Modeling Outputs

Figure 1b shows the GPS $S_{490}$max versus date for all stations, according to their longitude. The gray cells indicate data gaps, and the contours indicate the angle between the day-night terminator and the local magnetic field direction (i.e., the “alpha angle”) as functions of time and longitude.

The seasonal/longitudinal variations of scintillation occurrence exhibit patterns that are generally consistent with the alpha angle variations (in good agreement with Wu, 2015); TPU is a clear exception due to its location north of the anomaly peak. EPB occurrence is increased for places and times of the year for which the day-night terminator aligns well with the local magnetic field, to within ±15°. The scintillation occurrence measured at TPU is significantly reduced compared to the other stations, which mostly reside within ±15° latitude of the magnetic equator. Similar patterns are noticeable for the UHF data shown in Figure 1c, although the nonzero $S_{490}$max values are typically much larger than their GPS counterparts.

While the seasonal/longitudinal patterns can be well charaterized by the alpha angle, there exists a significant degree of daily variability in both the GPS and UHF data sets. This is particularly true for the Southeast Asian/Pacific sector, which exhibits a high degree of scintillation activity during the 2013 September and 2014 March equinoxes. In the South American/Atlantic sector, the strength of the GPS scintillation highly varies from September 2013 to April 2014 but nonetheless shows nonzero scintillation activity almost every day throughout this period; this lack of daily variability is more clear in the UHF data set, Figure 1c.

Figure 1d is similar to panels b and c but shows the daily maximum of the R-T growth rates calculated from the TIEGCM outputs for the respective GPS/UHF station meridians. The largest R-T growth rates tend to occur at times/places for which the alpha angle is small and show some similar characteristics to the observations shown in panels b and c. However, there are also some apparent disagreements between the model and observations: for example, the lack of a local minimum in the R-T growth rate during the December solstice for the Southeast Asian and Pacific sectors. One other observation worth noting from the TIEGCM R-T
growth rates is the significant daily variability for all station locations throughout the entire year, particularly for times during which significant scintillation activity was observed.

To examine more closely how well the TIEGCM R-T growth rate corresponds to the observed scintillation observations, Figure 2 shows the scintillation data and the TIEGCM growth rates during the period of 1 August to 31 October 2013 for the BKK station. Figures 2a and 2b show the hourly GPS and UHF S490, respectively, as functions of local time and date throughout this period; gray cells indicate data gaps. To highlight the daily variability in the observed scintillation levels, the yellow lines indicate the daily maximum values. Figure 2c shows the hourly maximum of the TIEGCM growth rate in the same format. Figure 2d shows the Kp index and the Fl0.7 solar flux during this period, that is, the two sources of daily variability within the TIEGCM (Carter, Yizengaw, et al., 2014).

Figure 2 shows a good level of agreement between the scintillation occurrence observed using GPS and UHF receivers. Scintillation activity is less common towards the beginning and end of the 3-month period shown and, when present, tends to begin after 19–20 LT and continues to local midnight (and sometimes later) in both data sets. In the UHF scintillation data, the levels exhibit a more binary/stronger variation from 1 day to the next compared to the GPS data. The GPS and UHF data sets share several common scintillation and non-scintillation days throughout this period, but of particular interest are the three non-scintillation periods of 2 October, 9–10 October, and 15–16 October. These periods correspond very well with three periods of very low R-T growth rate, as determined from the TIEGCM outputs. This agreement between model and observations can be understood by examining the Kp index variations in Figure 2d, which show geomagnetic storms during these three periods. These decreased R-T growth rates during geomagnetic storms echo the results of previous studies using the TIEGCM (e.g., Carter, Retterer, Yizengaw, Groves, et al., 2014; Carter, Retterer, Yizengaw, Wiens, et al., 2014; Carter, Yizengaw, et al., 2014; Carter et al., 2016). Interestingly, there are two more non-low-scintillation periods during 22–24 and 27–30 October with quite small Kp increases that do not appear to be well described by the TIEGCM R-T growth rate variations; these examples serve as a reminder that not all of the physical processes involved in EPB growth/suppression are accounted for in the modeling.

3.2. Quantifying “Forecast” Success Using Skill Scores

While the qualitative agreement between the modeled R-T growth rate and scintillation observations above is clear for three brief periods of elevated geomagnetic activity, what is not entirely clear is the overall performance of the model in a scintillation prediction/forecasting sense. To investigate this, the topic of forecast assessment is investigated using deterministic or “binary-event” (i.e., “yes”/“no”) forecasts and the simple forecast contingency table, Table 1, where a, b, c, and d represent the number of true positives, false positives, false negatives, and true negatives, respectively.

Figure 3 shows the measured (a) GPS and (b) UHF S490max values versus the corresponding TIEGCM R-T growth rate maxima for the 3-month period shown in Figure 2, that is, the yellow lines in Figure 2. Also shown as the dashed lines are the scintillation and growth rate thresholds of 0.2 and 0.2 × 10^{-3} s^{-1}, respectively. The digits in the four corners of Figures 3a and 3b indicate the number of points that fall within that particular quadrant, which effectively represent the values within the forecast contingency table, Table 1, that is, “true positives” in the top-right in Figures 3a and 3b, “false positives” in the bottom-right, “false negatives” in the top-left, and “true negatives” in the bottom left. As an indication of the “forecast” skill, (“forecast” is in quotes here because, technically, these values were obtained retrospectively using finalized solar and geomagnetic indices) the Heidke Skill Score (HSS) is shown along with the percentage of EPB days within the 3-month period and the percentage of days correctly “forecast” using these particular thresholds (i.e., the “percentage correct (PC)” value).

The HSS (Heidke, 1926) is defined by

$$HSS = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}$$

which exhibits a skill score of zero for a random forecast (e.g., flipping a coin) and a positive (negative)
value for higher (lower) skill than a random forecast. The HSS has a maximum value of 1 and a minimum value of $-1$. From the thresholds given for the GPS data in Figure 3a, $a = 41$, $b = 5$, $c = 29$, and $d = 17$, which gives a PC (i.e., "days correct") value of $(a + d)/(a + b + c + d) = 58/92 = 63.04\%$ and an HSS value of 0.26 from equation (2) above, which indicates some skill in the TIEGCM R-T growth rate over a random forecast. Similarly, the thresholds yield $a = 54$, $b = 9$, $c = 11$, and $d = 13$ for the UHF data in Figure 3b, giving $PC = 65/87 = 74.71\%$ and $HSS = 0.33$.

While some previous studies have used the HSS to indicate forecast skill (e.g., Carter, Retterer, Yizengaw, Groves, et al., 2014; Carter, Yizengaw, et al., 2014), it has not yet been established that this measure is appropriate for EPB forecasts. The period selected in Figure 3 is during peak EPB/scintillation activity. As such, one could rationalize that a comparison against a random forecast is not the most appropriate; this would certainly be true for stations in the South American sector that exhibited a low level of daily variability (see Figure 1). For some stations and times of the year, potentially a good measure of success is a comparison against a constant "yes" forecast throughout the peak EPB/scintillation period. In this case, the PC parameter compared to the percentage of observed EPB/scintillation days throughout the data set could be a good skill indicator. Again, for the case presented in Figure 3, the TIEGCM would be deemed to be performing well for both data sets: a PC value of 63\% versus 50\% occurrence for GPS and a PC value of 75\% versus 72\% occurrence for UHF. However, these measures of success could be highly susceptible to changes in the $S_{490}^{\text{max}}$ threshold and—in the case of this study—the R-T growth rate threshold.

### 3.2.1. Sensitivity of Various Skill Scores to Threshold Selection

While studies tend to agree on the level of $S4$ that constitutes a “scintillation event” (typically $S4 > 0.3$), there are reasons for why such a selection may not be appropriate for all stations and seasons. For instance, the BKK station used in this example is actually located rather close to the geomagnetic equator, where peak electron densities are lower than at the equatorial ionization anomaly peaks to the north and south. This is important because the $S4$ index is proportional to the background electron density (Whalen, 2009). Therefore, the selection of 0.2 as the $S_{490}^{\text{max}}$ threshold was chosen above for the BKK data. UHF signals, on the other hand, are quite sensitive to the presence of ionospheric irregularities and tend to show a strong binary-type response (see Figure 3b). As such, the selection of $S_{490}^{\text{max}}$ for UHF receivers can be rather loose and not have much impact on an analysis of model performance. Given that the choice of these thresholds is important for binary-event forecasts, it is worth investigating how threshold selection can impact the results of any model/technique assessment for EPB/scintillation forecasts.

The question of how the HSS varies with the selection of $S4$ and R-T growth rate thresholds in the BKK August–October 2013 data set is now explored. Figures 4a and 4b are the same as Figure 3 but are colored...
According to the HSS that results from that choice of $S_{490}^{max}$ and growth rate thresholds; the color bar is shown in panel (c). For example, using thresholds of $S_{490}^{max} = 0.25$ and $0.25 \times 10^{-3} \text{s}^{-1}$ for the growth rate, the HSS would be $\sim 0.4$ in both GPS and UHF data sets (as indicated by light blue cells in Figures 4a and 4b).

At a first glance, it is clear that the HSS strongly depends on which scintillation threshold is used and that this dependence can be quite different between different data types, in this case, co-located GPS and UHF scintillation receivers. Therefore, it is clear that a more systematic approach for threshold selection is needed.

Figure 4. $S_{490}^{max}$ versus R-T growth rate colored according to the skill score that results from choosing that $S_{490}^{max}$ and R-T growth rate threshold pair (i.e., where the dashed lines meet in Figure 3). Panels (a) and (b) show how the HSS varies with threshold selection for the GPS and UHF data, respectively. White cells indicate negative HSS values, and gray cells indicate regions outside of the data range. Panels (c) and (d) show the variation of the TSS$_{95}$ with threshold selection. In this case, gray cells indicate regions for which the confidence level lies below 95%. Panels (e) and (f) are the same as (c) and (d) but for the ORSS$_{95}$.

According to the HSS that results from that choice of $S_{490}^{max}$ and growth rate thresholds; the color bar is shown in panel (c). For example, using thresholds of $S_{490}^{max} = 0.25$ and $0.25 \times 10^{-3} \text{s}^{-1}$ for the growth rate, the HSS would be $\sim 0.4$ in both GPS and UHF data sets (as indicated by light blue cells in Figures 4a and 4b).

At a first glance, it is clear that the HSS strongly depends on which scintillation threshold is used and that this dependence can be quite different between different data types, in this case, co-located GPS and UHF scintillation receivers. Therefore, it is clear that a more systematic approach for threshold selection is needed.
At this point, it is appropriate to introduce the “True Skill Statistic” (TSS, also known as the Peirce Skill Score) (Woodcock, 1976) and the Odds Ratio Skill Score (ORSS) (Stephenson, 2000). The TSS is the difference between the hit rate and the false alarm rate; that is,

\[
TSS = H - F = \frac{a}{a + c} - \frac{b}{b + d} = \frac{ad - bc}{(a + c)(b + d)}. \tag{3}
\]

The TSS is commonly used in measuring forecast success (e.g., Barnes et al., 2016; Bloomfield et al., 2012; McGranaghan et al., 2018); however, it has not been customary to examine the confidence intervals on these TSS values. As discussed by Stephenson (2000), it is possible to perform confidence testing on the TSS using the TSS standard error:

\[
SE_{TSS} = \sqrt{\frac{H(1-H)}{n_H} + \frac{F(1-F)}{n_F}}, \tag{4}
\]

where \(n_H = a + c\) and \(n_F = b + d\). Given that the TSS is simply the hit rate minus the false alarm rate, one
can claim with 95% confidence that a given binary-event prediction technique exhibits skill only if the TSS $- 1.96 \ast SE_{\text{TSS}} > 0$. Significance testing is typically reliable, provided that the values in Table 1 are 5 or greater (Stephenson, 2000).

Figures 4c and 4d are the same as a and b but for the TSS with the 95% confidence interval applied, with the additional provision that all values in Table 1 are 5 or greater, $TSS_{95}$. Gray cells indicate thresholds for which these criteria were not met. Similar to the HSS plots in Figures 4a and 4b, the $TSS_{95}$ varies with the choice of $S4_{90\text{max}}$ and R-T growth rate thresholds, but the range of possible threshold choices is significantly reduced.

The odds ratio $\vartheta$ is defined as the odds of making a correct forecast to the odds of making an incorrect forecast (Stephenson, 2000);

$$\vartheta = \frac{ad}{bc}.$$  \hspace{1cm} (5)

The larger this value is, the higher the probability there is for a good forecast. The ORSS is calculated as

$$\text{ORSS} = \frac{\vartheta - 1}{\vartheta + 1},$$  \hspace{1cm} (6)

which has a range of $-1$ to 1, with $-1$ indicating a negative association, 0 indicating no skill, and 1 indicating perfect skill. Importantly, the ORSS can also have confidence intervals applied to it (Stephenson, 2000). The 95% confidence interval is determined for the ORSS using the “natural log odds”;

$$\log_e(\vartheta) > 0.$$  \hspace{1cm} (7)

The standard error, $SE_{\text{ORSS}}$, is given as $1/(n_h)^{1/2}$, where $n_h$ is the effective number of degrees of freedom; that is,

$$SE_{\text{ORSS}} = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)^{1/2}.$$  \hspace{1cm} (8)

From this, one can determine if the odds ratio is statistically significant using the criterion in equation (7);

$$\log_e(\vartheta) \pm 1.96SE_{\text{ORSS}} > 0.$$  \hspace{1cm} (9)

From this

$$\log_e(\vartheta) \pm \log_e(a^{1.96SE_{\text{ORSS}}}) > 0$$

$$\Rightarrow \log_e(a^{1.96SE_{\text{ORSS}}}) > 0.$$  \hspace{1cm} (10)

Taking the exponential of both sides gives

$$ae^{1.96SE_{\text{ORSS}}} > 1,$$  \hspace{1cm} (12)

which gives the upper and lower bounds of the 95% confidence levels to use in equation (6). Figures 4e and 4f are the same as Figures 4c and 4d but for the ORSS with the 95% confidence interval applied also with the condition that a, b, c, and d are 5 or greater, $ORSS_{95}$. While the values of the $ORSS_{95}$ in Figures 4e and 4f are generally larger compared to the $TSS_{95}$ values in Figures 4c and 4d, the range of statistically significant choices for $S4_{90\text{max}}$ and R-T growth rate thresholds is quite similar.

### 3.2.2. Sensitivity of Skill Scores to Data Set Type and Size

The data set considered in Figures 2–4 has been limited to 3 months (therefore consisting of 92 days/events), and with that quantity of data statistically significant $S4_{90\text{max}}$ and R-T growth rate thresholds have been shown to be achievable. Although, how many days/events does one need to achieve statistical significance?

To investigate this question, both BKK GPS and UHF 12-month data sets were broken up into time series of 30-, 50-, 60-, and 90-day sliding windows, and the skill scores were calculated for all combinations of
For each of the 30-, 50-, 60-, and 90-day periods, the thresholds yielding the best skill scores were selected, emulating natural researcher behavior (i.e., reporting the most favorable results). Figure 5 displays the results of this analysis, and how the HSS, TSS, and ORSS vary throughout the BKK GPS (a–c) and UHF (d–f) data sets, using sliding data windows of 30 (red), 50 (cyan), 60 (blue), and 90 days (black). To aid in interpreting Figure 5, the skill plotted for a given date was calculated using the data from 0 UT of that day ±n/2, where n = 30, 50, 60, and 90.

Interestingly, the period spanning November to January appears to have no statistically significant skill scores. This can be understood by reviewing the GPS S4 observations in Figure 1 for BKK. This period is

\[ S_{4\text{max}} \text{ and R-T growth rate thresholds.} \]

For each of the 30-, 50-, 60-, and 90-day periods, the thresholds yielding the best skill scores were selected, emulating natural researcher behavior (i.e., reporting the most favorable results). Figure 5 displays the results of this analysis, and how the HSS, TSS, and ORSS vary throughout the BKK GPS (a–c) and UHF (d–f) data sets, using sliding data windows of 30 (red), 50 (cyan), 60 (blue), and 90 days (black). To aid in interpreting Figure 5, the skill plotted for a given date was calculated using the data from 0 UT of that day ±n/2, where n = 30, 50, 60, and 90.

In the GPS data set (Figure 5a), it is clear that the HSS widely varies throughout the 1-year period and that the HSS can change significantly based on the size of the data set used, for example, mid-December HSS → 0 for the 90-day window (black) and HSS → 1 for the 30-day window (red). However, the range of TSS and ORSS values is more confined (Figures 5b and 5c), ranging between 0.2 and 0.6 across the full year for 50-, 60-, and 90-day data sets. Further, it becomes clear that there are very few TSS and ORSS values for the 30-day sliding window, which is an indication that 30 days is rarely enough data to achieve statistical significance.

Figure 6. Comparison of the ORSS values obtained for the TIEGCM R-T growth rate versus the persistence forecast (red) for the (a) 60- and (b) 90-day sliding window analyses using the BKK GPS data. The dashed lines indicate the boundaries of the 95% confidence level in the respective ORSS values, noting that when this confidence level fell below 0 the ORSS was not plotted. Panels (c) and (d) are the same as (a) and (b) but for the BKK UHF data set.
characterized by very little scintillation activity in the BKK GPS data and therefore exhibits almost no daily variability. As such, the 95% confidence limitations for the TSS and ORSS are effectively limiting the use of the TSS and ORSS to characterize forecast “success” during this period; the HSS calculation made no such limitations and actually resulted in quite high values for the 30-day data sets.

Finally, an important observation from the TSS and ORSS variations is that the 50-, 60-, and 90-day sliding windows all showed similar values, which implies that as long as the data set is large enough to achieve 95% confidence, the resulting skill score will be similar for data sets of different sizes. However, data sets larger than 90 days begin to include longer-term (e.g., seasonal and even solar activity) changes, which could impact the threshold selection(s) due to model/technique biases. Therefore, this analysis reveals that data set sizes between 50 and 90 days/events are sufficient choices for evaluating a model/technique’s ability to describe daily EPB occurrence variability.

Many of the observations discussed above can also be made for the UHF data set, Figures 5d–5f. That is, (1) the 30-day data sets rarely achieve the 95% confidence level, (2) the TSS and ORSS values remain fairly constant throughout the 1-year data set, and (3) the TSS and ORSS are relatively independent of the data set size/number of events, provided that the data set is between 50 and 90 days in order to achieve statistical significance. One difference is that the 90-day sliding window does achieve 95% confidence across the December solstice period, but this is most likely due to the size of the window capturing the leading and trailing edges of the scintillation seasons at the equinoxes. A significant result is that even though the UHF and GPS data sets are independent measures of scintillation activity, the skill scores reported for the GPS data in Figures 5b and 5c are comparable to those for the UHF data in Figures 5e and 5f.

3.2.3. Can Forecasting Models/Techniques Be Meaningfully Compared Using Skill Scores?
Figures 4 and 5 have demonstrated the potential benefits of using skill scores with statistical significance limitations. The TSS and ORSS skill scores show a good level of insensitivity to the selection of thresholds, data set size, and even data type. Next, the ability to use these skill scores to directly compare forecasting/prediction techniques/models is explored.

The ORSS is used to compare the TIEGCM R-T growth rate to the simple persistence forecast to determine if one prediction technique clearly outperforms the other. Similar to the analysis presented in Figure 5, the best ORSS value achieved for all possible thresholds is taken, with the added caveat for the persistence forecast that the predicted and observed S4 thresholds are the same; it would not be logical to allow an
observed scintillation event to be classified as, for example, $S_{490\text{max}} \geq 0.3$ but for the persistence forecast threshold to be $S_{490\text{max}} \geq 0.5$.

Figure 6 shows the (a) 60- and (b) 90-day sliding window ORSS$_{95}$ for the TIEGCM R-T growth rate compared to ORSS$_{95}$ for the persistence forecast (red) throughout the BKK GPS data set. The dashed lines indicate the boundaries of the 95% confidence interval, as given by equation (12). As with previous figures, the ORSS$_{95}$ is not plotted when the confidence intervals fall below 0. Figures 6c and 6d are same as Figures 6a and 6b but for the BKK UHF data set.

In Figures 6a and 6b, it can be seen that the persistence forecast and the TIEGCM R-T growth rate exhibit similar ORSS$_{95}$ values throughout the data set, with a couple of notable exceptions, namely, the period spanning mid-September until early November when the TIEGCM ORSS$_{95}$ confidence interval falls below 0 and the period from November until early January when only the persistence forecast had statistically significant ORSS values. While there are sections of the data set for which both the 60- and 90-day sliding windows reveal that the TIEGCM growth rate has a higher ORSS$_{95}$ than persistence, it is clear that the 95% confidence levels are generally larger than these differences. The same trends tend to apply to the ORSS$_{95}$ variations for the UHF data presented in Figures 6c and 6d.

One feature worth highlighting in Figures 6c and 6d is the lack of statistically significant skill score values for March–June for the 60-day sliding window persistence forecast (c) compared to the 90-day window (d). The lack of skill scores in this instance is not due to a lack of data but due to a lack of variability in the UHF data set, thus not satisfying the criterion that each element in Table 1 must be 5 or greater. This is demonstrated in Figure 7, which is the same as Figure 2 but for the March–April 2014 period. In particular, Figure 7b shows that there are almost consistently elevated scintillation levels observed throughout this 61-day period in the UHF data set and that the persistence forecast (which is shown by the orange dashed line) matches the observations on all but 8 days. While it may seem as though one can conclude that the persistence forecast performs well during this period, there is very little variability in the data set and thus does not constitute a valid test. This situation is analogous to choosing a 3-day period for which it rained and showing that a model “successfully” predicted rain for that entire period. While this fact might be true, it does not indicate that the model/technique has any real “skill” in terms of daily variability.

### 3.3. Using “Sufficiency” to Compare Models/Techniques

The results thus far have demonstrated the challenges and complications in using a single measure of skill/success in comparing the performance of two or more models in the context of scintillation prediction. One potential alternative uses the concept of “sufficiency” (Ehrendorfer & Murphy, 1988), which compares models using their conditional probabilities of forecasts given the observations, that is, $p(f|x)$, where $f$ and $x$ can only equal 0 or 1 for binary-event forecasts. In this framework, one model is deemed to produce higher quality forecasts than another when it can be shown to be “sufficient” for that other model.

More specifically, Ehrendorfer and Murphy (1988) defines that Model A is sufficient for Model B when a stochastic transformation $h(f^A|f^B)$ exists such that

$$h(f^A = 1|f^A = 1)p_{11}^A + h(f^A = 1|f^A = 0)(1 - p_{11}^A) = p_{11}^B,$$  \hspace{1cm} (13)

and

$$h(f^B = 1|f^A = 0)p_{01}^B + h(f^B = 1|f^A = 0)(1 - p_{01}^B) = p_{01}^B.$$  \hspace{1cm} (14)

where $0 \leq h(f^A|f^B) \leq 1$, $p_{11} = p(f = 1|x = 1)$ is the likelihood of a true positive, $p_{10} = p(f = 1|x = 0)$ is the likelihood of a false positive, and superscripts A and B denote Models A and B, respectively. In other words, if the results of Model B can be simulated by randomizing the results of Model A, then Model A is sufficient for Model B (Callies, 2000). If the stochastic transformation does not exist, then Model A is deemed to be insufficient for Model B; however, this scenario does not necessarily mean that Model B is sufficient for Model A unless the superscripts in the equations above can be switched and the stochastic transformation is found to exist. In the event that the stochastic transformation in both cases is found to not exist, then both models are deemed to be insufficient for each other (i.e., of comparable quality). For details concerning the concept of “sufficiency” the reader is directed to Ehrendorfer and Murphy (1988).
In this study, Ehrendorfer and Murphy (1988)'s "likelihood-base rate" interpretation of sufficiency is used in which the existence of a stochastic transformation is determined by substituting $u = h(f_B^1 | f_A^1)$ and $v = h(f_B^1 | f_A^0)$ in equations (13) and (14) to give

$$u p_{11}^A + v (1 - p_{11}^A) = p_{11}^B,$$

$$u p_{10}^A + v (1 - p_{10}^A) = p_{10}^B,$$

where $0 \leq u \leq 1$ and $0 \leq v \leq 1$ (see equations 11 and 12 in Ehrendorfer & Murphy, 1988). The limitations on $u$ and $v$ in order for Model A to be deemed sufficient for Model B lead to the following inequalities (see equations 32 and 33 in Ehrendorfer & Murphy, 1988);

$$\frac{(1 - p_{11}^A)}{(1 - p_{10}^A)} \leq \frac{p_{11}^B}{p_{10}^B} \leq \frac{p_{11}^A}{p_{10}^A},$$

$$\frac{(1 - p_{11}^A)}{(1 - p_{10}^A)} \leq \frac{(1 - p_{11}^B)}{(1 - p_{10}^B)} \leq \frac{p_{11}^A}{p_{10}^A}.$$

Essentially, these four inequalities give rise to three different scenarios in comparing Models A and B; (1) Model A is sufficient for Model B if they are satisfied, (2) Model B is sufficient for Model A if they are
satisfied with the superscripts reversed, and (3) both models are insufficient for each other if the inequalities are not satisfied in scenarios (1) and (2).

In this study, sufficiency is used to explore whether a meaningful comparison between the model R-T growth rate results of the TIEGCM and the persistence forecast can be made for all GPS and UHF data sets. It was demonstrated above that statistically meaningful ORSS values were generally achieved in data sets with more than 50 days/events, although the 90-day data sets more consistently yielded statistically significant skill scores (i.e., using statistically significant $S_{499,\text{max}}$ and R-T growth rate thresholds). As such, the following sufficiency analysis was conducted using the 90-day sliding window data sets.

For each 90-day sliding-window period throughout the full year of data for all stations, the “sufficiency” category was determined for the TIEGCM growth rate versus the persistence forecast. Figure 8 summarizes the results of the sufficiency analysis in a similar format to Figure 1. Figure 8a shows the map in Figure 1a for reference, and Figures 8b and 8c show the sufficiency results for the GPS and UHF data sets, respectively. Blue indicates that the TIEGCM R-T growth rate is sufficient for the persistence forecast, black indicates that the persistence forecast is sufficient for the TIEGCM R-T growth rate, and red indicates that both the TIEGCM R-T growth rate and the persistence forecast are insufficient for each other. Gray indicates that no statistically significant pair of R-T growth rate and S4 thresholds exists for either the persistence forecast or the TIEGCM R-T growth rate, and therefore, no sufficiency test could be meaningfully performed; see, for example, the period spanning October–December 2013 in the BKK GPS data set in Figures 6b and 8b.

Before describing the results of Figure 8, it is important to note that each colored cell represents an individual model-model comparison over a 90-day (i.e., ±45 days) data set, which could, hypothetically, be the focus of an individual scintillation model assessment study. As such, summarized in Figure 8 are the results of many comparisons between the TIEGCM R-T growth rate and the persistence forecast techniques.

It is clear from the sufficiency analysis in Figure 8 that there are several cases for which the persistence forecast is higher quality than the TIEGCM R-T growth rate (black cells), and vice versa (blue cells), although there are many more cases for which both models are of comparable quality (red cells). Generally, the sufficiency outcomes tend to agree between the GPS and UHF data sets, although not in all cases. For the BKK station, the TIEGCM R-T growth rates generally provide higher quality forecasts in the transition from the March equinox to the June solstice (in agreement with the results presented in Figure 6), whereas in the transition from the December solstice to the March equinox either the persistence model is higher quality, or the models are of comparable quality, depending on whether the GPS or the UHF data are considered. Similar patterns can be seen for the TPU and KWA stations. In the South American and Atlantic sectors, either the persistence forecast is higher quality, or it is of comparable quality to the TIEGCM R-T growth rate. A notable exception to this is the transition period from the March equinox to the June solstice for the ASI, DKR, and CVD UHF data sets for which the TIEGCM is shown to be higher quality, Figure 8c.

Perhaps one of the most significant results from this analysis is the large number of gray cells in Figure 8 (particularly the peak EPB season in South American/Atlantic sectors and the off-peak season in Southeast Asian/Pacific sectors), indicating a lack of statistically significant skill score values. These cells indicate a lack of variability in those 90-day data sets, which results in an inability to select observation and model thresholds, which then deem any model-model comparison unreliable.

4. Discussion

The current level of maturity of EPB prediction techniques now warrants detailed investigations into assessment metrics in order to guide the research community towards its goal of developing reliable day-to-day EPB/scintillation forecasts. In establishing assessment metrics, it is important that they are appropriate and meaningful given the characteristics of EPB occurrence and the wide range of existing models, techniques, and data types. Presented in this study was an analysis of 12 months of co-located GPS and UHF scintillation observations alongside the TIEGCM R-T growth rate, as an example prediction technique, across multiple longitude sectors. The use of both UHF and GPS data sets provided an opportunity to explore how much impact, if any, different observation types have on such model/technique assessment metrics. Further, the use of different skill scores in evaluating the TIEGCM R-T growth rate was examined,
including an analysis of the skill score sensitivities to threshold selection, data set type, and data set size. Finally, the concept of “sufficiency” was explored as a method to compare EPB/scintillation forecast techniques.

4.1. Defining “Success” in Predicting EPB Occurrence

At first, the concept of forecast/modeling “success” seems rather rudimentary, particularly in the context of binary-event forecasting; that is, it is ideal to have many correct “yes” and “no” forecasts and very low false positives and false negatives. For this reason, the “percentage correct” (PC) parameter has been used in the context of EPB forecasts (e.g., Anderson & Redmon, 2017; Anderson et al., 2004; Carter, Retterer, Yizengaw, Groves, et al., 2014; Carter, Retterer, Yizengaw, Wiens, et al., 2014; Carter, Yizengaw, et al., 2014). However, PC is highly susceptible to the selection of event-defining thresholds. In other words, one could hypothetically hedge their observations/predictions so that a seemingly high PC value is achieved. For example, in this study the selection of low $S_{4\text{max}}$ and R-T growth rate thresholds effectively increases the number of EPB days observed and modeled, resulting in a high PC value.

Some previous works have used the HSS as a measure of forecast/modeling skill (e.g., Carter, Retterer, Yizengaw, Groves, et al., 2014; Carter, Yizengaw, et al., 2014), as the HSS value indicates skill with respect to a random forecast. While for particular data sets, the occurrence of EPBs may appear to be effectively “random,” with an occurrence rate close to 50% across the data set (e.g., Figure 3a), the well-known seasonal/longitudinal variations in EPB occurrence indicate that it is not a random process (e.g., Aarons, 1993; Basu & Basu, 1985; Basu et al., 2002; Burke, Gentile, et al., 2004; Burke, Huang, et al., 2004; Carter, Retterer, Yizengaw, Groves, 2014; Carter et al., 2013; Dai et al., 2011; Nishioka et al., 2008; Retterer & Gentile, 2009; Su et al., 2008; Tsunoda, 1985; Yizengaw & Groves, 2018; Yizengaw et al., 2013). With this in mind, it might seem appropriate to use the Brier Skill Score (BSS) (Brier, 1950) to quantify model/forecast success in EPB occurrence studies, however, there are two caveats to using BSS; (1) an existing climatological model is needed to compare against, and (2) on the order of 100s of events are required in order to achieve a statistically significant result (Bradley et al., 2008). While climatological models for EPB/equatorial scintillation exist (e.g., Aveiro & Hysell, 2010; Cervera et al., 2001; Forte & Radicella, 2005; Hysell & Burcham, 2002; Secan et al., 1995), there is not yet an openly available community-wide standard model. Further, the requirement of such a large number of events/samples in order to achieve a statistically significant BSS value is not yet viable for ionospheric scintillation studies. Bloomfield et al. (2012) recommended using the TSS instead of the HSS in comparing solar flare forecasts, due to the HSS varying when confronted with different event/no-event sample ratios (Woodcock, 1976). Bloomfield et al. (2012) showed that by doubling the number of true positives and false negatives (i.e., holding the overall hit rate, $a/(a + c)$, constant) that TSS was invariant, whereas the HSS was not. McGranaghan et al. (2018) also made use of the TSS to evaluate the forecast skill in their high-latitude scintillation model that was built using machine learning. While the issue of statistical significance may not have been a problem in both of those studies due to the large data sets used, EPB forecast studies typically have far fewer data available and thus require the issue of statistical significance to be investigated.

The ample data used in the current analysis facilitated a detailed investigation into the issue of statistical significance when using the TSS and the ORSS. In particular, $TSS_{95}$ and $ORSS_{95}$ were used to explore the potentially contentious issue of threshold selection, which is necessary for binary-event forecast evaluation. Fully noting the typical behavior of researchers choosing the most favorable thresholds for their technique/model, it was found in this study that by applying 95% significance limitations on the threshold selection itself, and removing any human element from that selection, the range of thresholds available to choose was drastically reduced, Figure 4. This is important because the dependence of the GPS-derived S4 parameter on the background ionospheric plasma density (Whalen, 2009) means that one must dynamically choose reasonable thresholds for different data sets that observed different ionospheric conditions, for example, whether the equatorial anomaly peak/trough was observed or whether solar maximum/minimum conditions were observed. Of course, this rationale could be used to manually choose S4 thresholds for each data set as it is analyzed (e.g., as was done by Carter, Retterer, Yizengaw, Groves, et al., 2014), but forcing the threshold selection to yield statistically significant skill scores (and not simply the highest value) gives these types of analyses more rigor.
Aside from the threshold selection, the use of the observed S4 parameter to indicate EPB activity also varies within the literature. For example, Caton and Groves (2006), Redmon et al. (2010), and Anderson and Redmon (2017) have shown a preference for the “Total Mean Hourly S4” parameter, which is the sum of the hourly average S4 values during a given night, whereas Carter, Yizengaw, et al. (2014) used the daily maximum of the hourly averaged S4, and Carter, Retterer, Yizengaw, Wiens, et al. (2014) used the $S_{4\text{max}}$, as did the current study. While researchers can provide justifications for their threshold and parameter selections, the wider EPB research community would benefit from a more rigorous measure of success less susceptible to these aspects. Stephenson (2000) investigated the sensitivity of various skill scores to these, and other, aspects and showed that the ORSS was the least sensitive. The analysis reported here supports Stephenson (2000)'s findings, but it was found here that as long as the 95% confidence levels were applied, either the TSS or the ORSS could be used. Therefore, it is recommended from this analysis that researchers use either $TSS_{95}$ or $ORSS_{95}$ to ensure statistically sound threshold selections.

### 4.2. How Much Data Are Enough?

The question of how many days/events are required in order to quantify any forecast model/technique's skill was also explored in the current analysis. As mentioned in section 1, the number of days/events within previous EPB forecasting studies ranges from days (Carter et al., 2016; Hysell et al., 2019; Kelley & Retterer, 2008) to months (e.g., Anderson & Redmon, 2017; Anderson et al., 2004; Carter, Retterer, Yizengaw, Wiens, et al., 2014; Redmon et al., 2010). By using the $TSS_{95}$ and $ORSS_{95}$ values, it was found in this analysis that at least 50 days was typically required in order to achieve statistically significant skill scores (Figure 5). The implication of this finding is that quantifying any model/technique’s ability to capture EPB occurrence variability requires more than only a few events. Importantly, this significance test should be performed for each EPB prediction trial/investigation, as the minimum number of events could change between data sets and/or forecast models/techniques.

This study also imposed the lower limit of 5 on each element in the forecast contingency table, Table 1, meaning that any model-data comparison must have at least 10 incorrect forecasts; otherwise, the ORSS values become less reliable (Stephenson, 2000); a value of 0 for either b or c will mean that ORSS will be undefined (equation 5). The obvious caveat to this is that as a model/technique becomes very good, more events are going to be needed in order to properly quantify the skill in using that model/technique. Related to this is the issue of the duration of EPB peak and off-peak seasons for any given location(s). While statistically significant skill scores were achieved in this study with 50–90 days, it is clear that seasonal EPB occurrence variations begin to dominate over longer timescales (i.e., >3 months). This means that any model(s) under evaluation using longer data sets need both climatology and daily variability in EPB occurrence well-modeled.

### 4.3. Forecast Model/Technique Comparisons

Conducting performance comparisons with models and/or techniques is not a simple process. As argued throughout the meteorology literature, simple skill score comparisons do not necessarily make clear that one model outperforms, or is higher quality than, another (e.g., Ehrendorfer & Murphy, 1988; Murphy, 1993; Murphy & Ehrendorfer, 1987). The best way to make such comparisons is to conduct them on a 1-1 basis, that is, using the same data set/time interval. Given the current lack of availability/accessibility of equatorial ionospheric scintillation forecasting models and long-term data sets, conducting such 1-1 comparison studies is difficult. Further, just because a 1-1 comparison was done using one set of data, it does not mean that the results will apply employing a separate set of data. A perfect example of this is the prediction trial by Carter, Retterer, Yizengaw, Wiens, et al. (2014), which compared the results of the TIEGCM R-T growth rate versus the empirical model WBMOD and the persistence forecasts. It was shown in that analysis that the EPB season strongly influenced which model performed best. The TIEGCM R-T growth rate reproduced the EPB suppression days during peak EPB season due to geomagnetic activity, whereas both the WBMOD and the persistence forecast performed the best during the transition from the peak to off-peak season. Such considerations make it impossible to draw overarching conclusions as to the quality of one model/technique over another.

The persistence forecast has understandably become a useful forecast in the case of EPB occurrence due to its simplicity and accuracy. As can be seen in the South American sector, the persistence forecast rather
accurately describes the observations at both GPS and UHF frequencies, Figure 1. Of course, the major flaw of the persistence forecast is its inability to capture daily variability, which is the overarching goal of the EPB/scintillation prediction community at present. Therefore, the persistence forecast presents as a clear benchmark.

Another reason for why the persistence forecast is very useful in EPB occurrence prediction studies is the insight that it provides into the validity of the data set being used to test the daily variability within a given model/technique. Case studies limited to a few days certainly contribute significant value in terms of understanding why a given model/technique performs/does not perform well in particular cases/events. However, given the seasonal/longitudinal nature of EPB occurrence, it is important that the data set being used for forecast model/technique assessment contains enough day-to-day variability. The EPB occurrence observed from South American stations can tend to suffer from a lack of daily variability, as shown in Figure 1, and using a model to predict EPBs every night for a 3-month period when only two nights observed no EPBs is not particularly useful/insightful from a prediction assessment standpoint. Further, the current analysis focused on a period within the solar cycle for which the level of daily variability in EPB occurrence changed across locations and seasons. Under high/low solar activity conditions, the level of daily variability in EPB occurrence also changes. However, in applying the condition of statistical significance on the skill scores achieved by the persistence forecast, one can still ensure that the observation data set contains enough daily variability for any further analysis. In other words, during solar minimum when EPBs are rare, the persistence forecast will correctly predict no EPBs almost every day, but this would not yield either a TSS$_{95}$ or ORSS$_{95}$ value. As such, no threshold selection can be made, and therefore, no model-model comparison can be performed using that data set. The same is true for high solar activity during which EPBs are very common for specific seasons and/or locations, for example, the South American sector during the December solstice in Figure 1. Therefore, the persistence forecast was shown to provide a validity check on the observation data set used and to be a useful alternative EPB prediction model to compare against in the context of capturing daily EPB variability.

Given that several previous studies have utilized weeks-to-months of data to perform EPB prediction model/technique evaluations (e.g., Anderson et al., 2004; Anderson & Redmon, 2017; Carter, Retterer, Yizengaw, Groves, et al., 2014; Carter, Retterer, Yizengaw, Wiens, et al., 2014; Carter, Yizengaw, et al., 2014; Redmon et al., 2010; Retterer et al., 2005), the 12-month data sets used in this study were all broken up into 30-, 50-, 60-, and 90-day sliding window data sets to emulate the comparison conditions in those works but multiplied several-fold. The results for each day in Figures 5, 6, and 8 thus reflect the skill score/sufficiency outcome for the data period spanning that day ±$n$/2 days, where $n$ is 30, 50, 60, or 90 days. The analysis contained in Figure 6 revealed that the range of possible TSS and ORSS values within 95% confidence levels was generally too large to separate the performance of the TIEGCM R-T growth rate and the persistence forecast. So, an alternative model comparison technique that uses the concept of model “sufficiency” (Ehrendorfer & Murphy, 1988) was employed in this study.

As described in section 3, if the results of a Model B can be simulated by randomizing the results of Model A, then Model A is deemed to be sufficient for Model B; that is, that Model A produces higher quality forecasts than Model B. The sufficiency analysis revealed a few significant results. First, it showed that for the majority of cases, both the persistence and TIEGCM R-T growth rate forecasts were of comparable quality, red cells in Figure 8. The EPB seasons for which this was the case spans peak, transitional, and off-peak periods, although the TIEGCM R-T growth rate produced higher quality forecasts during some peak and transitional EPB occurrence periods, for example, March–June for ASI, BKK, and TPU. The persistence forecast tended to produce higher quality forecasts in the periods before and after peak EPB season in the South American and Atlantic sectors and for other sporadic periods. From these results, it is not clear as to whether the TIEGCM R-T growth rate outperforms the persistence forecast or vice versa. Ultimately, there are sections within the data for which one of these is true and other sections of the data for which the opposite is true. The fact that the results of this model-model comparison are so varied illustrates a valuable lesson in making careful conclusive statements from limited short-term studies, such as those emulated in this analysis.

The sufficiency analysis also demonstrated the value in using the persistence forecast as a benchmark because there were many 90-day data sets for which a statistically significant skill score (in this case, ORSS$_{95}$) could not be achieved, gray cells in Figure 8. Figure 7 showed a clear example of a 60-day period.
in the BKK UHF data set for which only 4 days observed little/no scintillation activity. Similar cases are common throughout the South American sector across 90-day periods, Figures 1b, 1c, and 8. While the lack of daily variability in the South American sector is quite well-known, the impact of this lack of variability on forecast model assessment has not been highlighted in the literature before and is a key finding that the community should be aware of as it strives to produce reliable day-to-day EPB occurrence forecasts.

A significant caveat to the use of “sufficiency” in EPB occurrence prediction assessment shown in this study is that simply shifting one’s focus from one data set to another can completely alter the outcome of the assessment. For example, if an analysis of 90 days of GPS scintillation data from BKK was used to validate the TIEGCM R-T growth rate in the transition period of March–June 2014, then the outcome would be that the TIEGCM outperforms the persistence forecast. However, if one chooses to analyze the period from December 2013 to March 2014, the outcome is completely reversed. As such, the sufficiency test, even with the statistical significance measures on threshold selection, does not provide a completely conclusive and robust model comparison in EPB occurrence prediction. The sufficiency test does provide meaningful information regarding the quality of the forecasts provided by a model compared to another, however in such comparisons, care must be taken to consider multiple datasets and seasons.

4.4. Tracking Future Progress in Predicting EPBs/Scintillations

The results of this analysis indicate that a single skill score that can be used to meaningfully compare models/techniques between separate data sets and data types, while exhibiting insensitivity to threshold selection and hedging, remains elusive. The results shown here demonstrate that comparisons of either HSS, TSS, or ORSS values between studies/data sets are not meaningful, as the issue of EPB prediction is more complicated with many considerations and nuances, as discussed above. This result echoes the many years of research on the topic of forecast assessment and verification in the meteorology literature (e.g., Ehrendorfer & Murphy, 1988; Murphy, 1993; Murphy & Ehrendorfer, 1987).

In order to track progress in EPB prediction models/techniques in future, direct access to the models in question will be required and 1-1 comparisons performed. For instance, if one were to evaluate the improvements/shortfalls of the newly developed Whole Atmosphere Community Climate Model with thermosphere and ionosphere extension (WACCM-X) (Liu, 2020; Liu et al., 2018) compared to one of its predecessors, the TIEGCM, in capturing daily variability of the R-T growth rate, then both models would need to be tested against the same collection of data, provided that the data exhibits enough daily variability. Even in such cases, the results of model-model comparisons are likely to depend on which period(s) and/or location(s) are considered, as revealed by Figure 8. Such comparisons will hopefully shed light on where future improvements/adaptations can be made, but there are many considerations and nuances, as demonstrated in this study. Therefore, tracking incremental progress in EPB prediction will continue to be a challenging task.

5. Summary and Conclusions

In this study, 12 months of co-located GPS and UHF scintillation data from the South American, Atlantic, Southeast Asian, and Pacific sectors were used to investigate methods and metrics for evaluating EPB occurrence forecast models/techniques, with a particular focus on day-to-day variability. As an example prediction technique, the R-T growth rates from the TIEGCM were used. Various commonly used skill scores were employed, and it was shown that calculating either the TSS or the ORSS, with their respective confidence levels, was a better choice compared to the HSS. Further, the use of confidence levels on the skill scores was shown to facilitate the selection of observation and modeling thresholds, which are necessary for binary-event forecasts. In this analysis, the confidence levels on the skill scores showed that data sets generally required between 50 and 90 days/events in order to achieve statistical significance. In terms of model-model comparisons, the persistence forecast was shown to be a useful benchmark. However, it was demonstrated that skill score values do not necessarily indicate which model produces higher quality forecasts. Therefore, the concept of “sufficiency” was explored as a method for conducting model-model comparisons, and this analysis provided some clear results. However, the use of sufficiency in such comparisons was shown to be not without its caveats and limitations, particularly how the model comparison outcome can be easily reversed by considering a different time interval/station location. An important result of this study was the use of the ORSS with its 95% confidence interval in determining whether a given observation period
contains enough daily variability within it. In this regard, not all periods and observation locations are equal when it comes to investigating daily variability; for example, the lack of daily variability in the South American sector makes it a rather poor location for such model/technique performance assessments against data. In striving to achieve reliable day-to-day EPB forecasts, care must be taken of the many nuances, not only in the detection and occurrence characteristics of EPBs but also in the assessment metrics and model comparison methods that are employed.

Data Availability Statement

The Kp index and F10.7 solar flux data were obtained from the OMNIWeb website (http://omniweb.gsfc.nasa.gov). The SCINDA data used in this analysis are available from the Zenodo data repository (doi:10.5281/zenodo.3872432).

Acknowledgments

This project was supported by the Australian Research Council Linkage Project scheme (project LP160100561). EY's research work was supported by the Aerospace Technical Investment Program and NSF (AGS-1848730) grant. The developers of the TIEGCM, the National Center for Atmospheric Research (NCAR) are gratefully acknowledged for their hard efforts in developing this model and making it available to the community.

References

Aaron, J. (1993). The longitudinal morphology of equatorial F-layer irregularities relevant to their occurrence. Space Science Reviews, 63, 209–243.
Abdu, M. A. (2012). Equatorial spread F/plasma bubble irregularities under storm time disturbance electric fields. Journal of Atmospheric and Solar-Terrestrial Physics, 75–76, 44–56.
Abdu, M. A., Bittencourt, J. A., & Batista, I. S. (1981). Magnetic declination control of the equatorial F region dynamo electric field development and spread F. Journal of Geophysical Research, 86, 11,443–11,446.
Anderson, D. N., & Redmon, R. J. (2017). Forecasting scintillation activity and equatorial spread-F. Space Weather, 15, 495–502. https://doi.org/10.1002/2016SW001554
Anderson, D. N., Reinisch, B., Valladares, C., Chau, J., & Veliz, O. (2004). Forecasting the occurrence of ionospheric scintillation activity in the equatorial ionosphere on a day-to-day basis. Journal of Atmospheric and Solar-Terrestrial Physics, 66, 1567–1572.
Anthes, R. A. (2011). Exploring earth’s atmosphere with radio occultation: Contributions to weather, climate and space weather. Atmospheric Measurement Techniques, 4(6), 1077–1103. https://doi.org/10.5194/amt-4-1077-2011
Anthes, R. A., Bernhardt, P. A., Chen, Y., Cucurull, L., Dymond, C. E., & Su, S. Y. (2004). Longitudinal variability of equatorial plasma bubbles observed by DMSP and ROCSAT-1. Journal of Geophysical Research, 109, A12S01. https://doi.org/10.1029/2004JA010583
Barker, B. A., Ram, S. T., Yizengaw, E., Pradipta, R., Retterer, J., Norman, R., et al. (2018). Unseasonal development of post-sunset EPBs in the African and Asian low-latitude regions. Geophysical Research Letters, 41, 8176–8184. https://doi.org/10.1002/2014GL062203
Barker, B. A., Yizengaw, E., Pradipta, R., Retterer, J. M., Groves, K., Valladares, C., et al. (2016). Global equatorial plasma bubble occurrence during the 2015 St. Patrick’s day storm. Journal of Geophysical Research: Space Physics, 121, 894–905. https://doi.org/10.1002/2015JA021914
Barker, B. A., Yizengaw, E., Retterer, J. M., Francis, M., Terkildsen, M., Marshall, R., et al. (2014). An analysis of the quiet-time day-to-day variability in the formation of post-sunset equatorial plasma bubbles in the Southeast Asian region. Journal of Geophysical Research: Space Physics, 119, 3206–3223. https://doi.org/10.1002/2013JA019570

CARTER ET AL. 21 of 23
Carter, B. A., Zhang, K., Norman, R., Kumar, V. V., & Kumar, S. (2013). On the occurrence of equatorial F-region irregularities during solar minimum using radio occultation measurements. Journal of Geophysical Research: Space Physics, 118, 892–904. https://doi.org/10.1029/2012JA017406

Caton, R., & Groves, K. (2006). Longitudinal correlation of equatorial ionospheric scintillation. Radio Science, 41, RS0232. https://doi.org/10.1029/2005RS003357

Cervera, M. A., Thomas, R. M., Groves, K. M., & Ramli, A. G. (2001). Ellfendy Validation of WBMOD in the Southeast Asian region. Radio Science, 36(5), 1559–1572. https://doi.org/10.1029/2000RS002520

Dao, E., Kelley, M. C., Roddy, P., Retterer, J., Ballenhi'n, J. O., de La Beaujardière, O., & Su, Y.-J. (2011). Longitudinal and seasonal dependence of nighttime equatorial plasma density irregularities during solar minimum detected on the C/NOFS satellite. Geophysical Research Letters, 38, L10104. https://doi.org/10.1029/2011GL047046

de La Beaujardière, O. (2004). C/NOFS: a mission to forecast scintillations. Journal of Atmospheric and Solar-Terrestrial Physics, 66(17), 1573–1591. https://doi.org/10.1016/j.jastp.2007.07.030

de Lima, G. R. T., Stephany, S., de 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

Dymond, K. F. (2012). Global observations of L band scintillation at solar minimum made by COSMIC. Radio Science, 47, RS0L18. https://doi.org/10.1029/2011RS004931

Ehrendorfer, M., & Murphy, A. H. (1988). Comparative evaluation of weather forecasting systems: Sufficiency, quality, and accuracy. Monthly Weather Review, 116(9), 1757–1770. https://doi.org/10.1175/1520-0493(1988)116<1757:CEOWFS>2.0.CO

Forté, B., & Radicella, S. M. (2005). Comparison of ionospheric scintillation models with experimental data for satellite navigation applications. Annals of Geophysics, 48(3), 505–514.

Gentile, L. C., Burke, W. J., & Rich, F. J. (2006). A global climatology for equatorial plasma bubbles in the topside ionosphere. Annales de Geophysique, 24, 163–172.

Goodman, J. L., & Kramer, L. (2001). Scintillation Effects on Space Shuttle GPS Data. Long Beach, CA, United States: ION National Technical Meeting. https://www.ion.org/publications/abstract.cfm?articleID=186

Groves, K. M., Basu, S., Weber, E. J., Smitham, M., Kuenzler, H., Valladares, C. E., Smitham, M., Kuenzler, H., & Rich, F. J. (2006). A global climatology for equatorial plasma bubbles in the topside ionosphere. Annales de Geophysique, 24, 163–172.

Hagan, M. E., Roble, R. G., & Hackney, J. (2001). Migrating thermospheric tides. Journal of Geophysical Research, 106(A7), 12,739–12,752.

Halford, A. J., Kellerman, A. C., Garcia-Sage, K., Klenzing, J., Carter, B. A., McGranaghan, R. M., et al. (2019). Application usability levels: A framework for tracking project product progress. Journal of Space Weather and Space Climate, 9, A34. https://doi.org/10.1051/swsc/2019030

Heeils, R. A., Lowell, J. K., & Spiro, R. W. (1982). A model of the high-latitude ionospheric convection pattern. Journal of Geophysical Research, 87(A8), 6339–6345. https://doi.org/10.1029/JA087iA08p06339

Heidke, P. (1926). Calculation of the success and goodness of strong wind forecasts in the storm warning service. The George-Anne, 8, 301–349.

Huang, C.-S. (2011). Occurrence of equatorial plasma bubbles during intense magnetic storms. International Journal of Geophysics, 2011, 401,858. https://doi.org/10.1155/2011/401858

Huang, C.-S. (2018). Effects of the post sunset vertical plasma drift on the generation of equatorial spread F. Progress in Earth and Planetary Science, 5, 3. https://doi.org/10.1186/s40645-017-0155-4

Ilyas, A., & Burcham, J. (2002). Long term studies of equatorial spread F using the JULIA radar at Jicamarca. Journal of Atmospheric and Solar-Terrestrial Physics, 64(12), 1531–1543. https://doi.org/10.1016/S1364-6826(02)00993-1

Ilyas, D. L., Millia, M. A., & Knuy, K. (2019). Radio beacon and radar assessment and forecasting of equatorial F region ionospheric stability. Journal of Geophysical Research: Space Physics, 124, 9511–9524. https://doi.org/10.1029/2019JA026991

Kelley, M. C., Larsen, M. P., LaHouc, C., & McClure, J. P. (1981). Gravity wave dependence of equatorial spread F. A case study. Journal of Geophysical Research, 86, 9087–9100. https://doi.org/10.1029/JA086iA11p09087

Kelley, M. C., Makela, J. J., de La Beaujardière, O., & Retterer, J. (2011). Convective ionospheric storms: A review. Reviews of Geophysics, 49, RG2003. https://doi.org/10.1029/2010RG000340

Kelley, M. C., & Retterer, J. (2008). First successful prediction of a convective equatorial ionospheric storm using solar wind parameters. Space Weather, 6, S08003. https://doi.org/10.1029/2007SW000381

Kikuchi, T., Hashimoto, K. K., & Nozaki, K. (2008). Penetration of magnetospheric electric field to the equator during a geomagnetic storm. Journal of Geophysical Research, 113, A06214. https://doi.org/10.1029/2007JA012628

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

Liu, H.-L. (2020). Day-to-day variability of prereversal enhancement in the vertical ion drift in response to large-scale forcing from the lower atmosphere. Space Weather, 18, e2015SW002334. https://doi.org/10.1029/2019SW002334

Liu, H.-L., Bardeen, C. G., Foster, B. T., Lauritzen, P., Liu, J., Lu, G., et al. (2018). Development and validation of the Whole Atmosphere Community Climate Model with thermosphere and ionosphere extension (WACCM-X 2.0). Journal of Advances in Modeling Earth Systems, 10, 401–402. https://doi.org/10.1175/2017MAMS01232

Maus, S., Macmillan, S., Chernova, T., Choi, S., Dater, D., Golovkov, V., et al. (2005). The 10th generation international geomagnetic reference field. Physics of the Earth and Planetary Interiors, 151, 320–322.

McGranaghan, R. M., Mannucci, A. J., Wilson, B., 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

Murphy, A. H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting. Monthly Weather Review, 121(11), 281–293. https://doi.org/10.1175/1520-0493(1993)121<0281:WGAFAC>2.0.CO

Murphy, A. H., & Ehrendorfer, M. (1987). On the relationship between the accuracy and value of forecasts in the cost-latitude ratio situation. Weather and Forecasting, 2(3), 243–251. https://doi.org/10.1175/1520-0434(1987)002<0243:OTRTA>2.0.CO

Nishioka, M., Saito, A., & Tsugawa, T. (2008). Occurrence characteristics of plasma bubble derived from global ground-based GPS receiver networks. Journal of Geophysical Research, 113, A05301. https://doi.org/10.1029/2007JA012605
CARTER ET AL.