Space Weather

RESEARCH ARTICLE
10.1029/2023SW003590

Special Section:
Machine Learning in Heliophysics

Validating Ionospheric Models Against Technologically Relevant Metrics

A. T. Chartier[1], J. Steele[1], G. Sugar[1,2], D. R. Themens[3], S. K. Vines[1], and J. D. Huba[4]

[1]Johns Hopkins Applied Physics Lab, Laurel, MD, USA, [2]Now at SpaceX, Hawthorne, CA, USA, [3]Department of Physics, University of New Brunswick, Fredericton, NB, Canada, [4]Syntek Technologies, Fairfax, VA, USA

Key Points:

- New, open access tools developed to validate ionospheric models
- Metrics are GPS position, HF communications, and critical frequency/NmF2
- Machine learning applied to "clean up" auto-scaled ionosonde data, removing 94% of badly scaled NmF2 values

Correspondence to:
A. T. Chartier,
alex.chartier@jhuapl.edu

Citation:
Chartier, A. T., Steele, J., Sugar, G., Themens, D. R., Vines, S. K., & Huba, J. D. (2023). Validating ionospheric models against technologically relevant metrics. *Space Weather*, *21*, e2023SW003590. https://doi.org/10.1029/2023SW003590

Received 8 JUN 2023
Accepted 13 NOV 2023

Abstract New, open access tools have been developed to validate ionospheric models in terms of technologically relevant metrics. These are ionospheric errors on GPS 3D position, HF ham radio communications, and peak F-region density. To demonstrate these tools, we have used output from Sami is Another Model of the Ionosphere (SAMI3) driven by high-latitude electric potentials derived from Active Magnetosphere and Planetary Electrodynamics Response Experiment, covering the first available month of operation using Iridium-NEXT data (March 2019). Output of this model is now available for visualization and download via https://sami3.jhuapl.edu. The GPS test indicates SAMI3 reduces ionospheric errors on 3D position solutions from 1.9 m with no model to 1.6 m on average (maximum error: 14.2 m without correction, 13.9 m with correction). SAMI3 predicts 55.5% of reported amateur radio links between 2–30 MHz and 500–2,000 km. Autoscaled and then machine learning "cleaned" Digisonde NmF2 data indicate a 1.0×10^{11} el. m^3 median positive bias in SAMI3 (equivalent to a 27% overestimation). The positive NmF2 bias is largest during the daytime, which may explain the relatively good performance in predicting HF links then. The underlying data sources and software used here are publicly available, so that interested groups may apply these tests to other models and time intervals.

Plain Language Summary Multiple research groups are developing models of the ionosphere to address effects on technology that depends on radio signal propagation. Here we present tests of these models that capture the model performance as it relates to some relevant applications. These are GPS position, long-distance HF communication and ionospheric critical frequency (which is proportional to the square root of the peak ionospheric density). All the data sets, testing code and model output used are made available in the public domain for reuse and development. For the test of critical frequency, we use machine learning to "clean up" the input data, removing errors introduced in the data generation process.

1. Introduction

With the development of coupled ionospheric models being pursued by multiple groups (e.g., H. L. Liu et al., 2018; Pham et al., 2022; Wang et al., 2014; Welling et al., 2015), there is a need for systematic and openly available validation tools. This investigation aims to provide such tools, and targets metrics relevant to technological applications, notably GPS positioning, HF communications links and ionospheric NmF2.

There have been previous studies addressing similar metrics. For example, Schreiner et al. (1999) tested the parameterized real time ionospheric model against radio occultation profiles and found a 13% root-mean-square error in foF2 (proportional to the square root of NmF2). Later, Scherliess et al. (2006) validated the Utah State Global Assimilation of Ionospheric Measurements Gauss-Markov against ionosonde NmF2 and altimeter-derived TEC, reporting a 20% mean absolute error against the Bear Lake dynasonde (located in Logan, Utah) and a ∼4 TECU bias against the altimeter TEC. More recently, Mitchell et al. (2017) demonstrated remarkable accuracy in predicting High-Frequency angles of arrival in the presence of traveling ionospheric disturbances, estimating 90% cumulative cone angle errors within 1.46° and 1.18° in two cases. There have also been validation studies testing first-principles models against public data, notably Shim et al. (2012) who used occultation-derived NmF2 and hmF2 and in situ electron density as the metrics. While they did not provide overall statistics, the NmF2 performance varied approximately from 1 to 3×10^{11} el. m^{-3} during quiet times, and about double that during storms.

For this effort, we aim to address several limitations of previous validation studies. First, all the validation tools, model output and data used are made public to allow for inspection, reuse and improvement. Second, we model

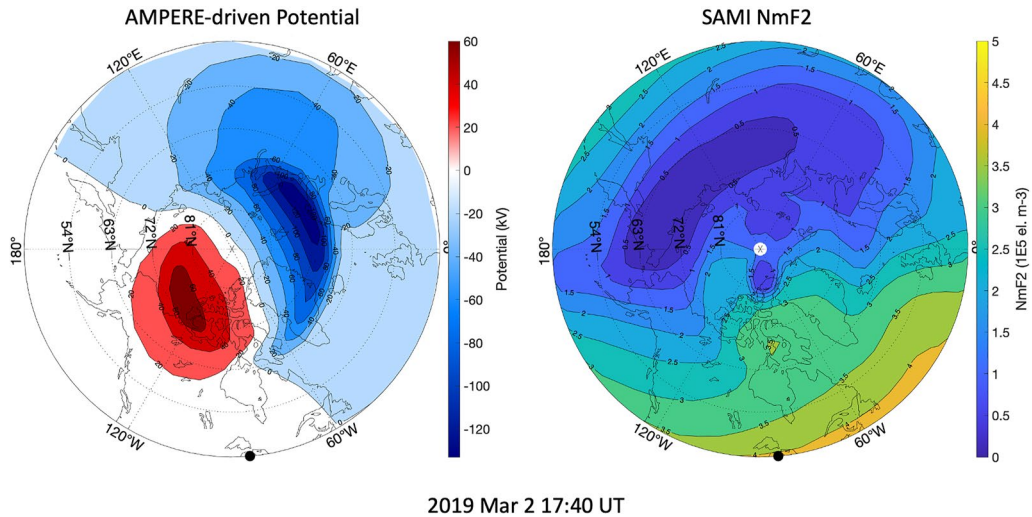

Figure 1. Left: The AMPERE-based potential, calculated using MIX with SAMI3 conductances. Right: The SAMI3 modeled NmF2 distribution, driven by the AMPERE-based potential.

ionospheric effects on technology to ease interpretation by the end user, quantifying ionospheric errors on GPS 3D position estimates and testing the model's ability to predict propagation paths used for HF communications. Third, we address a longstanding limitation of many ionosonde studies, by providing a machine learning algorithm to remove autoscaling errors.

The test model is a recent variant of Sami is Another Model of the Ionosphere (SAMI3, by Huba et al. (2000) and Huba and Joyce (2010)) as the test case. Here, SAMI3 is driven by a high-latitude potential based on Active Magnetosphere and Planetary Response Experiment (AMPERE, Anderson et al., 2000, 2014) field-aligned currents, using the Magnetosphere-Ionosphere-Coupling (MIX) approach of Merkin and Lyon (2010). The conductance model in the solver is provided by SAMI3. The neutral atmosphere is specified by the empirical thermosphere models NRLMSISE-00 (Picone et al., 2002) and HWM14 (Drob et al., 2015) and solar flux is quantified by the Flare Irradiance Spectral Model version 2 (Chamberlin et al., 2020). This arrangement was presented by Chartier et al. (2022). The AMPERE field-aligned current input is based on the first available month of Iridium-NEXT data (e.g., Califf et al., 2022), which is March 2019. An example of the output is shown in Figure 1.

This model is now being run routinely at the Applied Physics Laboratory, with the output available for visualization and download via https://sami3.jhuapl.edu. The output currently covers March, April, and May 2019, with the intention being to continue running it to cover the full AMPERE-NEXT data interval.

2. Method

Validation tools are designed to test model performance in three areas. First, ionospheric corrections to GPS single-frequency position estimates. Second, prediction of HF communications links. Third, prediction of F-layer critical frequency. The data coverage is shown in Figure 2.

As can be seen, the GPS provides good coverage at all latitudes, the ionosondes have good coverage at low-to-mid latitudes, and the Weak Signal Propagation Reporter (WSPR) link midpoints are concentrated primarily in Europe, North America, and Australia. Details of each validation technique are presented in the following three subsections.

2.1. GPS Position

The ionosphere constitutes a source of error on GPS position estimates as it produces a group delay and phase advance on the signals. At the L1 frequency (1,575.42 MHz), a column density of 1 TEC unit (TECU) produces a delay equivalent to a range error of 0.163 m (e.g., Komjathy et al., 2005). However it is not straightforward to interpret ionospheric TEC maps in terms of position error experienced by GPS users, as the position is calculated

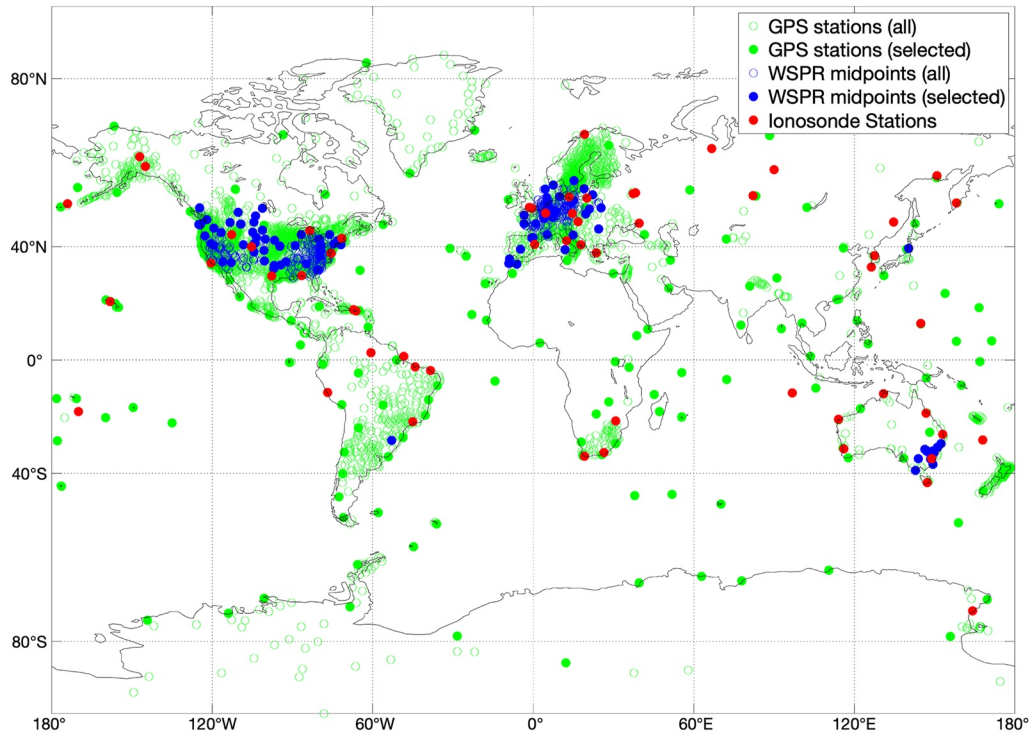

Figure 2. GPS, Weak Signal Propagation Reporter (WSPR) and ionosonde data used in this study. Open circles indicate datapoints discarded for computational efficiency, while filled circles are used in the analysis. The WSPR midpoints are shown for one representative time interval (0–1 UT on 15 March 2019).

using signals that take slant paths from multiple satellites to the receiver. Effects such as Geometric Dilution of Precision caused by the number and location of satellites in view must be taken into account. To that end, we estimate receiver positions from synthetic single-frequency pseudorange data for satellites actually in view at a set of test receivers at the specified times, with ionospheric delays generated using the TEC observed by dual-frequency receivers (processed by Massachusetts Institute of Technology group, Vierinen et al., 2016). The procedure is as follows:

1. Generate a list of 150 maximally separated GPS receiver sites from the ~4,000 dual-frequency sites available
2. Select those sites with >3 GPS satellites in view at the relevant time
3. Create synthetic pseudorange observations, using the geometric satellite-receiver distances plus the observed ionospheric dual-frequency TEC delays
4. Solve for the 3D position of the receivers from the synthetic pseudorange data (this is the "raw" position) —details below
5. Subtract the ionospheric delays predicted by SAMI3 and solve again (this is the "corrected" position)
6. Calculate position errors as the geometric distance between the true receiver position and the estimated position

Note that all non-ionospheric GPS range errors are neglected for this test.

The receiver position is calculated in a two-step process, as follows. First, the approximate (~100 m accuracy) position is calculated using a standard linear time difference of arrival (TDoA) multilateration approach. Then the linear TDoA position estimate is used as a starting point for the minimization described in Equation 1.

$$\arg.\ \min.\ f(rx) = \sum \left((rx - tx_i)^2 - pseudorange^2 \right)^2 \theta_i \quad (1)$$

The result is an estimate of the receiver position, rx. The sum is over the known GPS satellites in view, tx_i. *pseudorange* is calculated as described above. All positions and distances are specified in cartesian coordinates, in

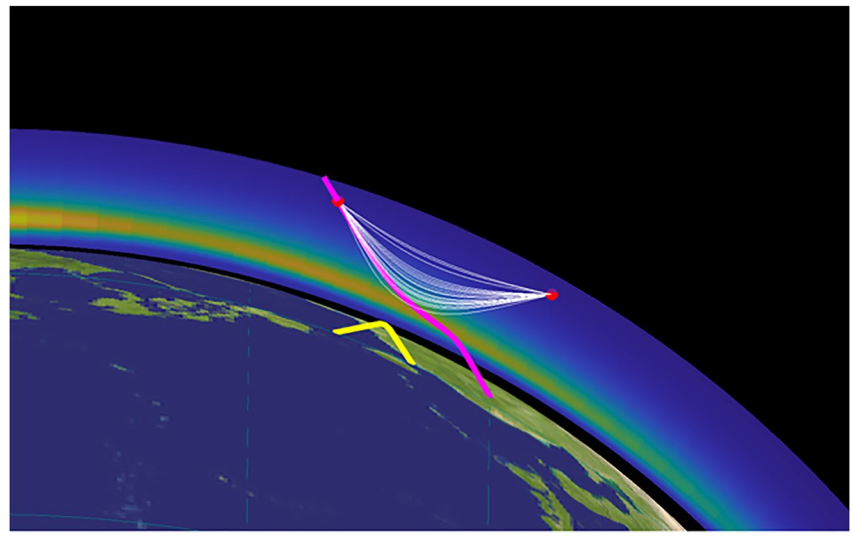

Figure 3. The new 3D homing algorithm enables ground-to-ground (yellow), ground-to-space (magenta), and space-to-space (white) point-to-point raytracing. Satellites indicated in red.

meters and θ_i is the elevation angle of the satellites as viewed from the receiver, specified in degrees. The elevation angle is included so that the solution favors satellites closer to zenith, which are likely to suffer smaller errors.

Note that multiple positioning algorithms exist, for example, "precise point positioning" (e.g., Bisnath & Gao, 2009) now in widespread use, often combined with the "real time kinematic" base station technique to provide rapid centimeter-level accuracy (e.g., Wübbena et al., 2005). However our simple approach still provides a means of quantifying many important aspects of the problem, notably the number and position of GPS satellites visible to a receiver on the ground. Precise point positioning algorithms can still benefit from accurate ionospheric solutions, for example, Li et al. (2022).

2.2. HF Communications

The ability of the ionosphere to reflect radio signals back down to Earth is the central property that led to its discovery. Several important technological applications make use of this property, notably geolocation, skywave over-the-horizon radar and long-distance communications. In recent years, radio amateurs (or "hams") have begun probing HF propagation paths (e.g., Frissell et al., 2014) using the WSPR protocol designed by Joseph Taylor, a nobel laureate and ham. A substantial database of radio links has been established, that we have used to test propagation paths in ionospheric models. From the full database of WSPR links (totaling 3 GB for March 2019) we select links between 2 and 30 MHz and between 500 and 2,000 km great circle distance in order to isolate likely single-hop ionospheric propagation. Then, we "declump" the data set according to their midpoints. An example of this process is shown in Figure 2. For 0 UT on 15 March 2019, there are 737 suitable links. The declumping algorithm selects 200 well-separated links at each timestep, based on the distance between their midpoints. Those links with the largest separation from their 10 nearest neighbors are included.

To evaluate whether an observed link is expected to exist, a three-dimensional point-to-point raytracing homing algorithm has been produced. This algorithm relies on the Provision of High-frequency Raytracing Laboratory for Propagation Studies (PHaRLAP) engine of Cervera and Harris (2014), though the homing algorithm could equally be applied to other raytracers. The new algorithm supports ground-to-ground, ground-to-space, and space-to-space homing. For the sake of computational efficiency, we have neglected magneto-ionic mode splitting from the results presented here. Figure 3 shows ground-to-ground, ground-to-space, and space-to-space links identified by the new homing algorithm.

The homing proceeds in two steps. First, a "global" search is performed where rays are shot across multiple elevations and azimuths. Then the optimization begins from the ray that passed closest to the target. For the WSPR case, we confine the "global" search azimuths close to the great-circle bearing between the transmitter and

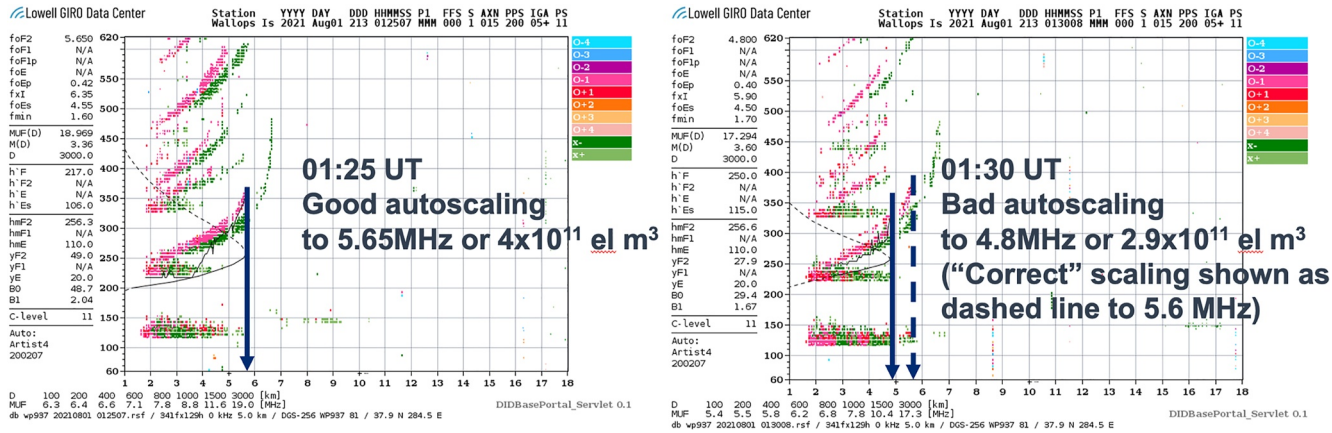

Figure 4. Example ionograms of accurate (left) and inaccurate (right) ARTIST autoscaling. These ionograms were taken about 10 min apart on 1 August 2021 at Wallops Island, and while the data appear very similar, ARTIST extracted foF2 values differing by 0.85 MHz. Upon inspection of the right panel it is clear that ARTIST misinterpreted the O-mode returns above 5 MHz between 325 and 380 km virtual range, resulting in an underestimate of foF2.

receiver. MATLAB's "fminsearch" is used to optimize the rays' initial elevations and azimuths until one hits the target. By default, a ray is considered to have hit the target when it passes within 5 km of it, though this parameter is user-defined. We confine the WSPR analysis to use signals propagating along great-circle distances of between 500–2,000 km and 2–30 MHz to eliminate non-ionospheric and multi-hop propagation effects. The test evaluates whether ionospheric propagation paths can be found through the SAMI3 model, for the links reported in the WSPR database.

2.3. Critical Frequency/Peak Density

F2-layer critical frequency (foF2), proportional to the peak electron density (NmF2), is one of the most important ionospheric parameters, both in terms of basic research and technological applications (e.g., Gardiner-Garden et al., 2019; Z. Liu et al., 2019). The global Digisonde network (Reinisch, 1995) presents a tremendous resource of foF2 observations as well as other parameters. Sixty One digisonde stations are available for model validation in this study.

The Global Ionospheric Radio Observatory (GIRO) network provides good spatial and temporal coverage over low and mid latitudes, with a few stations available at high latitudes (Reinisch & Galkin, 2011). However there have been concerns over the reliability of auto-scaled data products (e.g., Stankov et al., 2023; Themens et al., 2022). To address that limitation, we developed a machine learning classifier to flag unreliable datapoints. The "truth" data used in training here is a large set of 34,968 manually scaled ionograms produced by D. R. Themens. Manually scaled ionogram data are widely accepted to be among the most accurate forms of peak electron density measurement (e.g., Gilbert & Smith, 1988). One advantage of the new machine learning classifier is that it relies only on the autoscaled output, and does not require access to the raw ionograms or depend on other geophysical parameters.

Figure 4 shows examples of accurate and inaccurate autoscaling. On the left, the Automatic Real-Time Ionogram Scaler with True height (ARTIST, described by Galkin et al. (2008)) finds the maximum ordinary-mode return correctly, while on the right it misses the peak by almost 1 MHz. We note that multiple versions of ARTIST exist—this investigation is limited to the publicly available output from GIRO, in which the ARTIST version varies by station. There have been many attempts to detect and improve the performance, one of the most widely known being QualScan (McNamara, 2006). However, the QualScan software is not available for public use, test and validation.

To address this issue, we used a manually scaled set of ionograms from 33 stations covering all latitude bands. Using this data set as a ground truth, it is possible to test the accuracy of the ARTIST confidence scores. We define accurate autoscaled foF2 values as those within 0.5 MHz of the manual scalings, and find 12% of the data fail that test. Figure 5 shows that both accurate and inaccurate foF2 values are most often given confidence scores

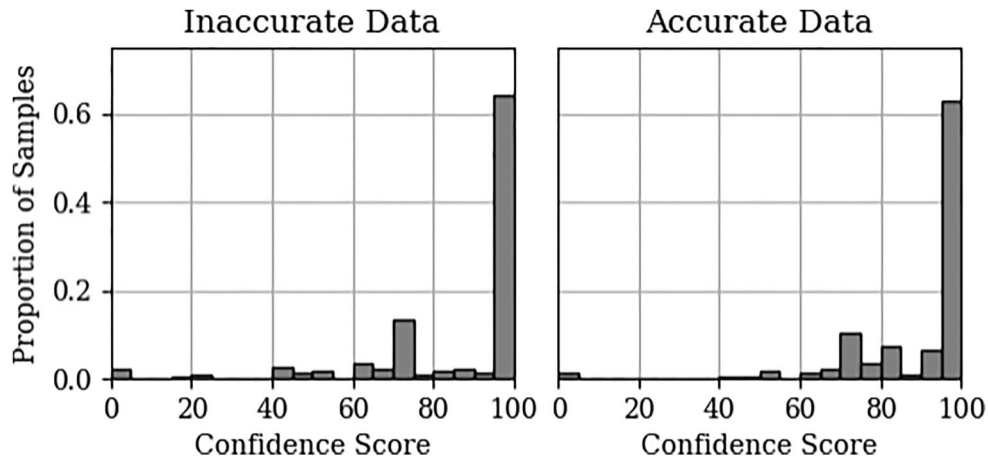

Figure 5. Weighted histograms of ARTIST-provided confidence scores for inaccurate (left) and accurate (right) ARTIST-extracted foF2 values. Note that the proportion of inaccurate data with high (>95) confidence scores is roughly the same as that for the accurate data (64.3% and 62.7%, respectively). "Accurate" data are defined as those lying within 0.5 MHz of the manually scaled values.

of 100. In fact, if one discards all data with a score <100, only 36.8% of inaccurate foF2 values will be correctly classified as such, and 45.3% of accurate foF2 values will be discarded.

Clearly a better means of distinguishing between accurate and inaccurate autoscaled foF2 data is required. To this end, we developed a machine-learning classifier based on time series of autoscaled ionogram parameters. The input data are the ARTIST-extracted foF2 and hmF2 across a two-hour window, plus the ARTIST confidence scores. We require that there be at least one ARTIST-extracted foF2 and hmF2 value before and one after the one in question. The classifier decides whether the autoscaled foF2 value is accurate or inaccurate, based on statistical properties of data in the two-hour window (see Appendix A for details).

6,362 ionograms did not pass the data availability requirement. The remaining ionograms were divided into 20,024 for training, and 8,582 for testing the classifier. A random forest classifier was trained with the intention of maximizing rejection of inaccurate data (true positive), without substantially increasing rejection of accurate data (false positive) as compared to selection of data with ARTIST confidence scores of 100. In Appendix A, we provide a description of the classifier development, including the full 27-element feature vector and hyperparameters used. The result was a classifier with a similar false positive (rejection of accurate data) performance on the test data (46.7% rejected by the ML, vs. 45.3% of accurate data lost due to <100 confidence score), while having a far better performance correctly flagging inaccurate foF2 values (93.7%, vs. just 36.8% of bad data flagged with <100 confidence score). With proper rejection of inaccurate data, autoscaled data from the global Digisonde network provides a valuable resource for ionospheric model validation.

3. Results

The period of study for all aspects of the validation is 2–31 March 2019. 1 March 2019 was removed because the model initialized from a nominal starting condition and requires 24 hr to approach a more realistic ionosphere.

3.1. GPS Validation

The GPS position analysis was carried out on the 150 stations identified in Figure 2. Results were divided into 10° and 1-hr bins, as shown in Figure 6.

This test indicates use of SAMI3 could produce a modest improvement in GPS 3D position accuracy at midlatitudes (24.5% improvement) and during the daytime (31.3% improvement). However, at low and high latitudes the model has little or no positive effect on the solution (10.1% and −1.2%, respectively). Likewise at night the improvement is only 14.9%. We note that the position errors are typically small in the March 2019 test period, and

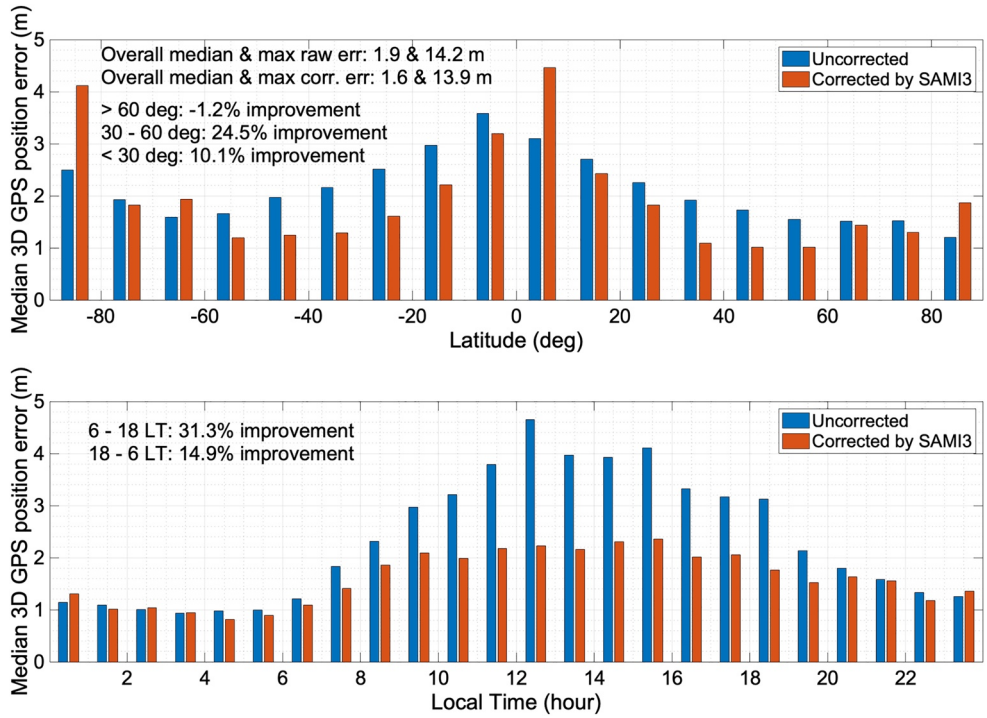

Figure 6. Statistics of ionospheric 3D GPS positioning error estimates, based on observed dual-frequency TEC data from 150 stations distributed worldwide. The algorithm accounts only for ionospheric contributions to position error.

so the uncertainty in our reference "truth" dual-frequency TEC (typically 3–5 TECU or 0.5–0.8 m) may account for a substantial fraction of the total error.

3.2. WSPRnet Validation

The WSPRnet validation tests whether the model predicts a radio propagation path on a specified frequency between two points, in cases where a corresponding link has been reported by radio amateurs and logged in the database. This test is "one-sided" by definition—links that were not made are not considered, and therefore a perfectly reflecting mirror ionosphere model would receive a perfect score while a model that reflects no signals would receive zero. A constraint to use links between 500–2,000 km is designed to prioritize likely single-hop ionospheric propagation over other modes. This constraint has the unfortunate side effect of limiting geographical coverage primarily to Europe, North America and Australia. Figure 7 shows an example of the WSPRnet validation from 15 March 2019 at 22:00 UT. The links shown are drawn from those logged by the WSPR protocol and reported in the public database. Links are colored green when the homing raytracing is able to find a corresponding path through SAMI3, while links that do not close through the model are colored red.

The WSPR validation was applied to hourly output from SAMI3 for the month of March 2019, with up to 200 links selected for analysis at each timestep (depending on availability). The results of the test are shown in Figure 8, binned in terms of latitude and local time of the link midpoints (the assumed reflection points for single-hop propagation). Of 128,258 test links, 71,141 closed through SAMI3, representing a 55.5% success rate. The latitudinal coverage is limited, so it is not possible to identify a clear trend in latitudinal performance. The local time analysis shows better performance during the day, with 69.7% success between 6 and 18 LT versus 38.5% success between 18 and 6 LT.

3.3. Ionosonde Validation

The ionosonde validation was carried out using autoscaled NmF2 data from all 61 available Digisonde stations worldwide. Our machine learning classifier was used to flag and remove likely inaccurate (>0.5 MHz error) autoscaled foF2 data before conversion to NmF2. This removes ~40% of all data-points available. The remaining

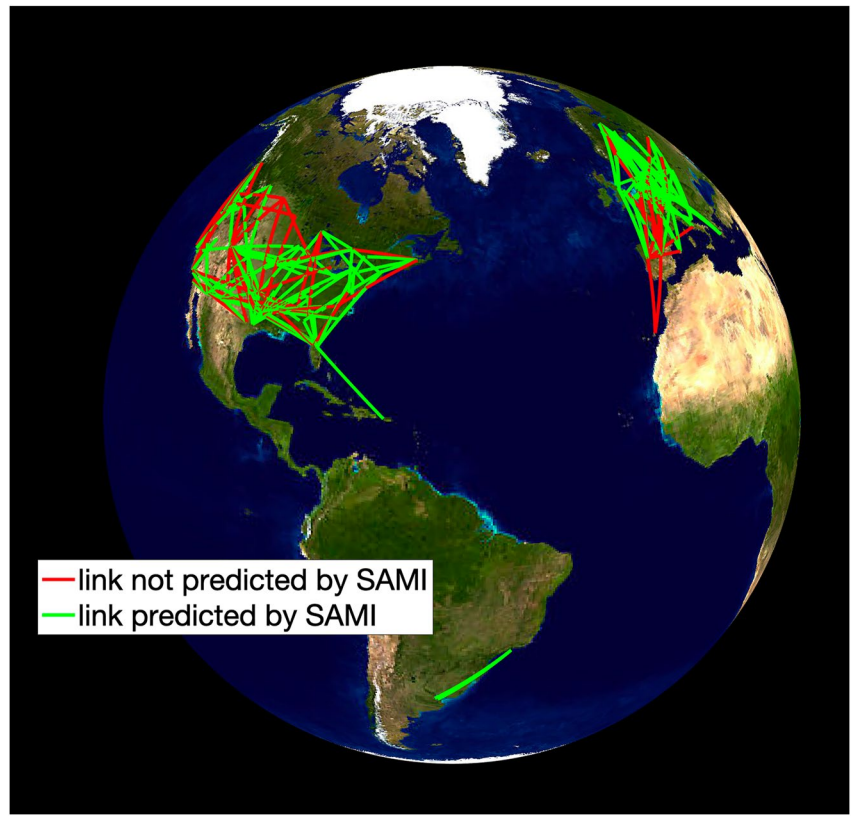

Figure 7. Weak Signal Propagation Reporter validation on 15 March 2019 at 2,200 (based on 200 site subsample).

data are compared against SAMI3 NmF2, which is linearly interpolated to the location of the ionosonde stations. Model values within 10 min of the observation time are considered acceptable for comparison, although the temporal discrepancy is generally less than 5 min (the model output is stored at 10 min cadence). SAMI3 NmF2 errors are analyzed using a binned box and whiskers approach, with 10° latitude and 1 hr local time divisions. Results are shown in Figure 9.

The analysis indicates a positive bias of NmF2 in all latitude and local time sectors, with a median value of 1.0×10^{11} el. m^{-3}. For reference, the median observed value of NmF2 is 3.7×10^{11} el. m^{-3}. The bias is largest in the local daytime (1.4×10^{11} el. m^{-3} between 6 and 18 LT vs. 0.5×10^{11} el. m^{-3} outside those hours) and at low latitudes (2.2×10^{11} el. m^{-3} below 30° vs. 0.7×10^{11} el. m^{-3} above 30°). Larger errors are expected at low latitudes and during the day because NmF2 values are generally higher there, but a systematic bias is not necessarily expected. Minimum and maximum errors of -1.36 and $+1.10 \times 10^{12}$ el. m^{-3} are observed, with full details in Table 1.

4. Discussion

We have developed new, publicly available validation tools to assess ionospheric models against operationally relevant metrics using open data. The model output tested in this case is SAMI3 driven by AMPERE NEXT-derived high latitude potential solutions, for the month of March 2019 (the first month for which AMPERE NEXT output is available). The metrics are correction of GPS 3D position estimates, closure of reported HF radio links through the model ionosphere, and prediction of ionospheric peak density (NmF2). These validation metrics have different attributes: GPS provides the best spatial coverage and is relevant to positioning applications, but is not easily interpreted in terms of effects on HF propagation. The WSPR data set is directly relevant to HF communications, but is mostly limited to North America and Europe due to the requirement for 500–2,000 km links (to test likely single-hop propagation). Ionosondes provide an excellent indication of the ionospheric critical frequency that could be relevant to over-the-horizon radar and geolocation applications, and we were able to retrieve data from

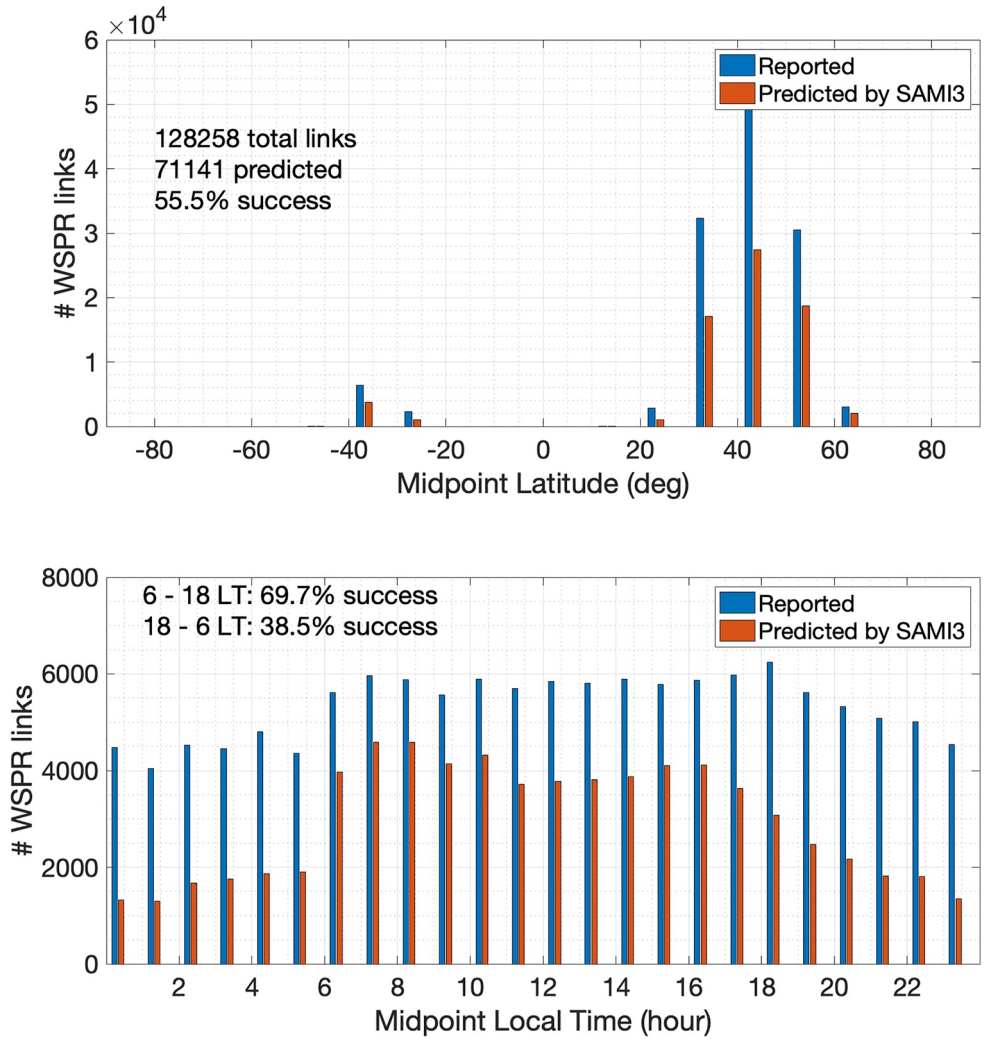

Figure 8. Weak Signal Propagation Reporter radio links covering 500–2,000 km great circle paths in the 2–30 MHz band, from March 2019. Up to 200 links per hour are selected, with counts shown in blue. SAMI3-predicted links are shown in red.

over 60 Digisondes worldwide for this study. Nevertheless coverage is limited at high northern and southern latitudes, and in many longitude sectors. Given that this is the first application of these metrics, it is not possible to state whether the model performs "well" or "badly" in absolute terms. However, certain trends can be identified and compared against prior expectations.

The modeled GPS position correction offers only a modest improvement from 1.9 to 1.6 m on average (maximum error: 14.2 m without correction, 13.9 m with correction), with a few cases found where the model makes the solution worse. This might be expected given our test interval (March 2019) is geomagnetically quiet with low solar activity, and so the initial ionospheric error is small. The input test data (dual-frequency TEC) may contain biases on the order of 3–5 TECU, equivalent to 0.5–0.8 m of range error. Unsurprisingly, the performance improvement is best at midlatitudes and during the daytime, where the ionosphere is relatively smooth and dense. At low latitudes the presence of steep density gradients may be more challenging to model, while the low densities at high latitudes and at night make it harder to gain any improvement. We note that a previous study by Rakipi et al. (2015) obtained a very large ionospheric correction averaging 43 m in the vertical direction at a midlatitude station in Albania (41°N, 20°E), using Klobuchar (1987)'s ionospheric model. This correction was much larger than their tropospheric correction of ∼3–4 m, and is much larger even than our uncorrected midlatitude ionospheric errors. The effect of ionospheric scintillation on position estimation (addressed by Jerez et al. (2019)) is a separate topic to bulk TEC time delay, and is not addressed by SAMI3 or by Klobuchar's model.

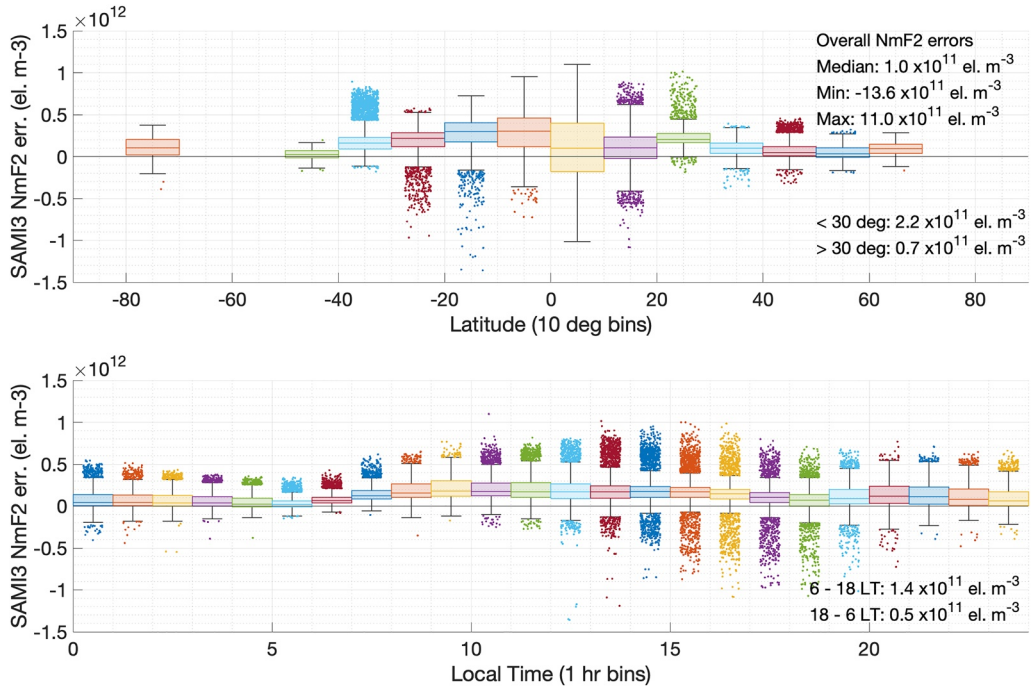

Figure 9. Box and whiskers analysis of SAMI3 NmF2 errors for March 2019, as compared against ML-classified autoscaled Digisonde data. Dots show outliers >1.5× the interquartile range. Above plot shows a breakdown in geographic latitude and below shows a breakdown in local time.

The WSPR validation is intended to test whether the model can predict HF radio communications links logged in the public database. In general, the propagation path of these signals is not known. For this test, the search space is restricted to paths between 500–2,000 km, to prioritize likely single hop propagation over other modes. The test is inherently "one-sided" in the sense that only observed "good" links are tested—a faulty model might predict ubiquitous propagation and thereby receive a perfect score. Additionally, the test as currently implemented does not account for absorption, both because the true directional performance of the amateur radio systems is unknown and because this version of SAMI3 lacks a D-region. Therefore it is important to consider the WSPR test alongside other data, in particular the ionosonde NmF2 test. Nevertheless, we find that SAMI3 can account for over half of the 128,258 reported links used (71,141, or 55.5%), with the performance much better during the day than at night (69.7% vs. 38.5%).

The test against autoscaled ionosonde NmF2 provides a more conventional test, assessing the model's performance in predicting one the most important ionospheric parameters. The peak density, NmF2 (derived from the observed critical frequency, foF2), is the most reliable parameter observed by ionosondes because it does not rely on any form of inversion, unlike other parameters such as hmF2. However, autoscaling errors can occur and these typically result in underestimation caused by missing the true peak of the ionogram either due to attenuation from strong derivative absorption near the peak in the ionogram or interference and restricted propagation bands that, if large enough, can cause some versions of ARTIST to stop scaling earlier in the trace. Consistent with Themens et al. (2022), our comparison against manually scaled data indicates that the ARTIST-provided "confidence score" is of limited value in identifying badly scaled data, with 73.2% of "bad" (>0.5-MHz from the manually scaled value) foF2 points assigned a confidence score of 100. We developed a ML algorithm that relies only on time series of autoscaled parameters as input, and rejects all but 6.3% of the "bad" data in the test set. The attrition rate is high in both cases, with 46.7% of "good" datapoints eliminated by the ML algorithm, versus 45.3% eliminated by rejection of data with <100 confidence score. The performance of SAMI3 in predicting NmF2 was assessed using the ML-cleaned data.

Table 1
Minimum and Maximum Errors of SAMI3 NmF2

Time	Station ID	Lat (°N)	Lon (°E)	Error (10^{12} el. m^{-3})
16 March 2019 14:10	CS31K	−12.2	96.8	−1.36
30 March 2019 05:40	BVJ03	2.8	−60.7	1.10

Overall, in the month of March 2019, we find the model has a 1.0×10^{11} el. m^3 median positive bias, with errors ranging from -13.6 to $+11.0 \times 10^{11}$ el. m^3. The positive bias is largest during the day (1.4×10^{11} el. m^{-3} between 6 and 18 LT vs. 0.5×10^{11} el. m^{-3} outside those hours) and at low latitudes (2.2×10^{11} el. m^{-3} below 30° vs. 0.7×10^{11} el. m^{-3} above 30°). This may explain why the model had greater success in predicting WSPR links during the day than at night, since larger values of NmF2 are more conducive to ground-to-ground HF propagation and our test is inherently one-sided (i.e., it does not account for "false positive" propagation predictions). We note that other validation studies of SAMI3 covering different periods have not uncovered the same positive bias, for example, Chou et al. (2023) found SAMI3 had a normalized skill score of 0.95–1 for mean TEC error in two storms, indicating the model was essentially unbiased. Further study would be required to determine whether this is due to the chosen periods of study, or to the choice of validation parameter (TEC vs. NmF2).

Finally, it is noted that SAMI3 uses the empirical thermosphere models NRLMSISE-00 (Picone et al., 2002) and HWM14 (Drob et al., 2015) in this study. This also adds an uncertainty in the SAMI3 results given the importance of thermospheric densities and winds in determining ionospheric conditions. Future validation efforts will use a physics-based model of the thermosphere such as TIE-GCM to drive SAMI3 (see Huba et al. (2017) for details); this is expected to improve model performance.

5. Conclusions

New tools have been developed to validate models of ionospheric electron density. To demonstrate these tools, we have analyzed the SAMI3 model run for March 2019 with AMPERE-derived high-latitude potential solutions. The model output is available for visualization and download from https://sami3.jhuapl.edu and will be updated for more recent periods as possible. The underlying software for all these tools is made available with the intention that interested parties can apply them to other models.

Three new validation tools are presented to test performance in terms of GPS 3D position, HF radio communications, and NmF2. For our test case, the results indicate that SAMI3 reduces ionospheric errors on GPS 3D position solutions from 1.9 to 1.6 m on average, with more pronounced improvement found at low and mid latitudes and during the day where electron densities (and therefore ionospheric errors on position) are higher). The HF propagation test showed that SAMI3 predicts 55.5% of reported links between 2–30 MHz and 500–2,000 km, again performing better during the day. The test against NmF2 indicated that SAMI3 has a 1.0×10^{11} el. m^3 median positive bias (27% of the median observed value of 3.7×10^{11} el. m^{-3}). The bias is largest during the daytime, which may explain the relatively good performance in predicting HF links then.

Appendix A: Development of the Ionogram Classifier

We began our analysis with a data set of 34,968 manually scaled ionograms from 33 stations for our analysis. These stations are shown in Figure A1.

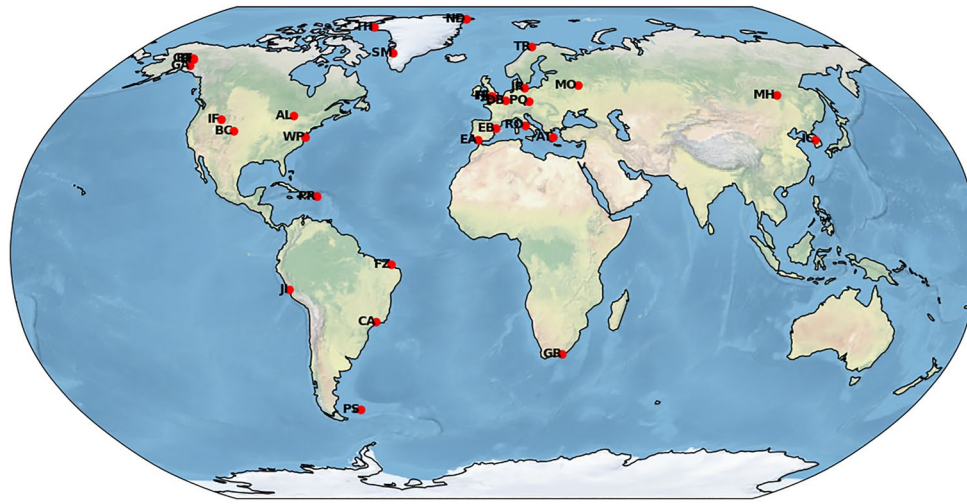

Figure A1. Ionosonde stations used to train the ML classifier.

For each ionogram, we generated a 27 element feature vector that contains information about the current ionogram and pre- and post-measurement windows containing ionogram data from 1 hr before and after the current ionogram, respectively. A description of each feature is given in Table A1. Some of the features are parameters generated from least squares regression line fitting to the pre- and post-measurement windows. Therefore, we required that there be at least two ionograms with ARTIST-extracted foF2 and hmF2 values in both windows. 6,362 ionograms did not pass this requirement, resulting in a data set size of 28,606 ionograms used for training and testing the classifier.

Table A1
Features Used in Building the Ionogram Classifier

Feature	Value
1	foF2 value of measurement
2	Pre-measurement window foF2 mean
3	Pre-measurement window foF2 standard deviation
4	Post-measurement window foF2 mean
5	Post-measurement window foF2 standard deviation
6	Centered window foF2 least squares regression line slope
7	Centered window foF2 least squares regression line Pearson correlation coefficient
8	Centered window foF2 least squares regression line p-value
9	Centered window foF2 least squares regression line slope standard error
10	Pre-measurement window foF2 least squares regression line slope
11	Pre-measurement window foF2 least squares regression line Pearson correlation coefficient
12	Pre-measurement window foF2 least squares regression line p-value
13	Pre-measurement window foF2 least squares regression line standard error
14	Post-measurement window foF2 least squares regression line slope
15	Post-measurement window foF2 least squares regression line Pearson correlation coefficient
16	Post-measurement window foF2 least squares regression line p-value

Table A1
Continued

Feature	Value
17	Post-measurement window foF2 least squares regression line standard error
18	hmf value of measurement
19	Centered window hmf standard deviation
20	Centered window hmf mean
21	Centered window hmf least squares regression line slope
22	Centered window hmf least squares regression line Pearson correlation coefficient
23	Centered window hmf least squares regression line p-value
24	Centered window hmf least squares regression line standard error
25	Confidence score of ARTIST measurement
26	Pre-measurement window confidence score minimum
27	Post-measurement window confidence score minimum

We used the scikit-learn library to train multiple random forest classifiers to classify ARTIST autoscaling as either accurate or inaccurate, where accurate autoscaled foF2 values are within 0.5 MHz of the manual scaled foF2 values and inaccurate values are not (Pedregosa et al., 2011). Classifiers were trained to maximize the F-beta (F_β) score, where F_β is the weighted harmonic mean of precision and recall:

$$F_\beta = \left(1 + \beta^2\right) \frac{\text{precision} \times \text{recall}}{\beta^2\,\text{precision} + \text{recall}} \tag{A1}$$

β is a parameter that controls whether the classifier favors precision or recall. If $\beta < 1$, the classifier gives more weight to precision (i.e., fewer false positives), while if $\beta > 1$, the classifier gives more weight to recall (i.e., fewer false negatives). Because the classifier is trained to detect inaccurate ARTIST foF2 estimates, false positives are defined as when the classifier predicts an inaccurate label when the value is actually accurate, while false negatives are defined as when the classifier predicts an accurate label when the value is actually inaccurate. Therefore as β increases, the classifier becomes more strict because the cost of a false negative (i.e., an inaccurate value being classified as accurate) increases compared to the cost of a false positive (i.e., an accurate value being classified as inaccurate). Classifiers were trained across a grid search of the hyperparameters identified in Table A2.

We trained the classifiers using β values between 0.25 and 16 as follows. First, we split the 28,606 ionogram data set into a 20,024 ionogram training set and a 8,582 ionogram test set. We then performed a grid search to determine the best hyperparameter combinations from Table 1 to use for a random forest classifier for a given β using 3-fold cross validation over the training set. The resulting optimal classifiers' true positive and false positive rates on the test set for a given β are shown in the last two rows of Table A3. As expected, both the false and true positive rates increase with β. If the classifier is being used as a filter to discard inaccurate values, a high β classifier

Table A2
Classifier Hyperparameters Used in Grid Search on the Scikit-Learn Random Forest Classifier

Hyperparameter and values used in grid search	
max_depth	3, 5, 10, ∞
min_samples_split	2, 4, 8, 13, 16, 21
min_samples_leaf	1, 10, 25, 50
max_features	1, 3, 5, 10, 20
class_weight	5, 50, 100, 150, 200, 300
n_estimators	4, 10, 50, 100, 200, 300

Note. The class_weight values are the relative weights given to samples identified as inaccurate.

Table A3
Optimal Hyperparameters Found With a Grid Search Over Values in Table A2, and the Corresponding True and False Positive Rates for the Test Set, for Different Beta Values

Hyperparameter	$\beta = 0.25$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$	$\beta = 4$	$\beta = 8$	$\beta = 16$
max_depth	3	3	∞	5	3	3	5
min_samples_split	8	16	13	16	2	2	4
min_samples_leaf	1	50	25	25	50	25	25
max_features	3	20	10	20	10	5	1
class_weight	5	5	200	50	50	150	300
n_estimators	10	4	4	10	4	100	100
True positive rate	0.28	0.49	0.743	0.911	0.937	1	1
False positive rate	0.035	0.099	0.159	0.358	0.467	0.825	0.972

Note. We selected the $\beta = 4$ classifier due to its high success rate in identifying inaccurate foF2 estimates (93.7%) and relatively low rate of misclassifying accurate estimates as inaccurate (46.7%).

will result in a small set of filtered values with high confidence that the filtered values are accurate, while a low β classifier will result in a large set of filtered values with a lower confidence that the filtered values are accurate. This is shown in Figure A2.

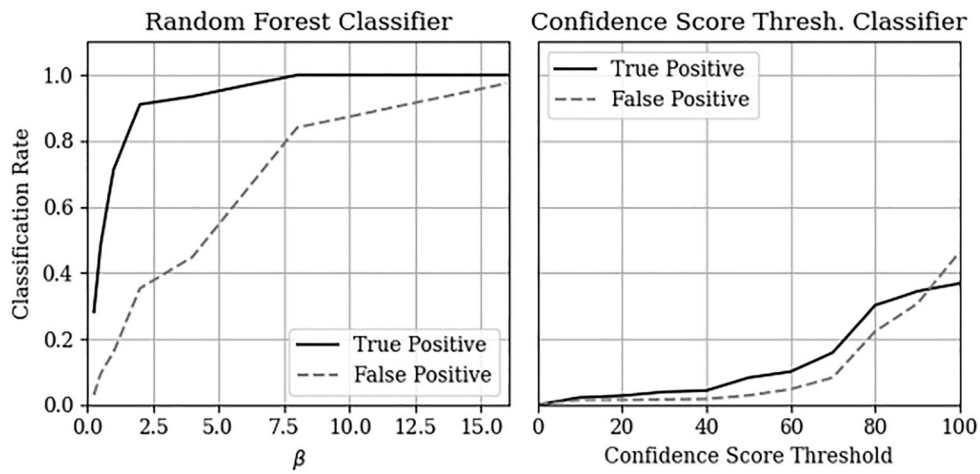

Figure A2. The true positive (solid) and false positive (dashed) rates for the optimized random forest classifiers as a function of β (left) and the simple ARTIST confidence score threshold classifier as a function of threshold (right). The random forest classifiers' performances are calculated using only the test set, while the confidence score performances are calculated using the full set.

Data Availability Statement

The code and ionogram classifiers were produced by Chartier and Sugar (2023). The results were provided by Chartier (2023). The MIT-TEC data are obtained from the MIT Haystack Madrigal data set by Rideout (2023a) while the Millstone Hill ISR data are provided by Rideout (2023b). Autoscaled ionogram parameters can be obtained from the Global Ionospheric Radio Observatory (Reinisch & Galkin, 2023). The PHaRLAP raytracer is available from Cervera (2023). SAMI3 model output is provided by Huba et al. (2023) and can be downloaded from https://sami3.jhuapl.edu. AMPERE data are provided by Vines (2023).

Acknowledgments
This paper uses ionospheric data from the USAF NEXION Digisonde network, the NEXION Program Manager is Annette Parsons. This product contains or makes use of IARPA data from the HFGeo program. The IARPA Program Manager is Dr. Torreon Creekmore. GS and ATC acknowledge support of NASA Heliophysics Grant 80NSSC21K1557. DRT acknowledges the support of Canadian Space Agency Grant 21SUSTCHAI.

References

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*(9), 3017–3025. https://doi.org/10.1002/2014gl059941

Anderson, B. J., Takahashi, K., & Toth, B. A. (2000). Sensing global Birkeland currents with Iridium® engineering magnetometer data. *Geophysical Research Letters*, *27*(24), 4045–4048. https://doi.org/10.1029/2000gl000094

Bisnath, S., & Gao, Y. (2009). Precise point positioning. *GPS World*, *20*(4), 43–50.

Califf, S., Alken, P., Chulliat, A., Anderson, B., Rock, K., Vines, S., et al. (2022). Investigation of geomagnetic reference models based on the Iridium® constellation. *Earth Planets and Space*, *74*(1), 37. https://doi.org/10.1186/s40623-022-01574-w

Cervera, M. A. (2023). Pharlap - Provision of high-frequency raytracing laboratory for propagation studies [Software]. Retrieved from https://www.dst.defence.gov.au/our-technologies/pharlap-provision-high-frequency-raytracing-laboratory-propagation-studies

Cervera, M. A., & Harris, T. J. (2014). Modelling ionospheric disturbance features in quasi-vertically incident ionograms using 3D magneto-ionic raytracing and atmospheric gravity waves. *Journal of Geophysical Research: Space Physics*, *119*(1), 431–440. https://doi.org/10.1002/2013JA019247

Chamberlin, P. C., Eparvier, F. G., Knoer, V., Leise, H., Pankratz, A., Snow, M., et al. (2020). The flare irradiance spectral model-version 2 (FISM2). *Space Weather*, *18*(12), e2020SW002588. https://doi.org/10.1029/2020sw002588

Chartier, A. T. (2023). Data for "validating ionospheric models against technologically relevant metrics" [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.8370338

Chartier, A. T., Huba, J. D., Sitaram, D. P., Merkin, V. G., Anderson, B. J., & Vines, S. K. (2022). High-latitude electrodynamics specified in Sami3 using ampere field-aligned currents. *Space Weather*, *20*(1), e2021SW002890. https://doi.org/10.1029/2021sw002890

Chartier, A. T., & Sugar, G. (2023). Code for "validating ionospheric models against technologically relevant metrics. *Zenodo*. https://doi.org/10.5281/zenodo.8381247

Chou, M.-Y., Yue, J., Wang, J., Huba, J. D., El Alaoui, M., Kuznetsova, M. M., et al. (2023). Validation of ionospheric modeled TEC in the equatorial ionosphere during the 2013 March and 2021 November geomagnetic storms. *Space Weather*, *21*(6), e2023SW003480. https://doi.org/10.1029/2023SW003480

Drob, D. P., Emmert, J. T., Meriwether, J. W., Makela, J. J., Doornbos, E., Conde, M., et al. (2015). An update to the Horizontal Wind Model (HWM): The quiet time thermosphere. *Earth and Space Science*, *2*(7), 301–319. https://doi.org/10.1002/2014EA000089

Frissell, N. A., Miller, E. S., Kaeppler, S. R., Ceglia, F., Pascoe, D., Sinanis, N., et al. (2014). Ionospheric sounding using real-time amateur radio reporting networks. *Space Weather*, *12*, 651–656. https://doi.org/10.1002/2014SW001132

Galkin, I. A., Khmyrov, G. M., Kozlov, A. V., Reinisch, B. W., Huang, X., & Paznukhov, V. V. (2008). The artist 5. In *AIP conference proceedings* (Vol. 974, No. (1), pp. 150–159). American Institute of Physics.

Gardiner-Garden, R., Cervera, M., Debnam, R., Harris, T., Heitmann, A., Holdsworth, D., et al. (2019). A description of the elevation sensitive oblique incidence sounder experiment (ELOISE). *Advances in Space Research*, *64*(10), 1887–1914. https://doi.org/10.1016/j.asr.2019.03.036

Gilbert, J. D., & Smith, R. W. (1988). A comparison between the automatic ionogram scaling system ARTIST and the standard manual method. *Radio Science*, *23*(06), 968–974. https://doi.org/10.1029/rs023i006p00968

Huba, J. D., Chartier, A. T., & Steele, J. (2023). Sami3 data in netCDF format (2019-Mar) (version 01) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7887963

Huba, J. D., & Joyce, G. (2010). Global modeling of equatorial plasma bubbles. *Geophysical Research Letters*, *37*(17), L17104. https://doi.org/10.1029/2010gl044281

Huba, J. D., Joyce, G., & Fedder, J. A. (2000). Sami2 is another model of the ionosphere (Sami2): A new low-latitude ionosphere model. *Journal of Geophysical Research*, *105*(A10), 23035–23053. https://doi.org/10.1029/2000ja000035

Huba, J. D., Maute, A., & Crowley, G. (2017). Sami3_ICON: Model of the ionosphere/plasmasphere system. *Space Science Reviews*, *212*(1–2), 731–742. https://doi.org/10.1007/s11214-017-0415-z

Jerez, G. O., Alves, D. B. M., & Tachibana, V. M. (2019). Multivariate analysis of combined GPS/GLONASS point positioning performance in Brazilian regions under different ionospheric conditions. *Journal of Atmospheric and Solar-Terrestrial Physics*, *187*, 1–9. https://doi.org/10.1016/j.jastp.2019.03.003

Klobuchar, J. A. (1987). Ionospheric time-delay algorithm for single-frequencyGPS users. *IEEE Transactions on Aerospace And Electronic Systems*, *23*(3), 325–331. https://doi.org/10.1109/taes.1987.310829

Komjathy, A., Sparks, L., Wilson, B. D., & Mannucci, A. J. (2005). Automated daily processing of more than 1000 ground-based GPS receivers for studying intense ionospheric storms. *Radio Science*, *40*(06), 1–11. https://doi.org/10.1029/2005rs003279

Li, P., Cui, B., Hu, J., Liu, X., Zhang, X., Ge, M., & Schuh, H. (2022). PPP-RTK considering the ionosphere uncertainty with cross-validation. *Satellite Navigation*, *3*(1), 10. https://doi.org/10.1186/s43020-022-00071-5

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*(2), 381–402. https://doi.org/10.1002/2017ms001232

Liu, Z., Fang, H., Hoque, M. M., Weng, L., Yang, S., & Gao, Z. (2019). A new empirical model of NmF2 based on CHAMP, GRACE, and COSMIC radio occultation. *Remote Sensing*, *11*(11), 1386. https://doi.org/10.3390/rs11111386

McNamara, L. F. (2006). Quality figures and error bars for autoscaled digisonde vertical incidence ionograms. *Radio Science*, *41*(04), 1–16. https://doi.org/10.1029/2005rs003440

Merkin, V. G., & Lyon, J. G. (2010). Effects of the low-latitude ionospheric boundary condition on the global magnetosphere. *Journal of Geophysical Research*, *115*(A10), A10202. https://doi.org/10.1029/2010ja015461

Mitchell, C. N., Rankov, N. R., B````ust, G. S., Miller, E., Gaussiran, T., Calfas, R., et al. (2017). Ionospheric data assimilation applied to HF geolocation in the presence of traveling ionospheric disturbances. *Radio Science*, *52*(7), 829–840. https://doi.org/10.1002/2016RS006187

Pedregosa, C. 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*, 2825–2830.

Pham, K. H., Zhang, B., Sorathia, K., Dang, T., Wang, W., Merkin, V., et al. (2022). Thermospheric density perturbations produced by traveling atmospheric disturbances during August 2005 storm. *Journal of Geophysical Research: Space Physics*, *127*(2), e2021JA030071. https://doi.org/10.1029/2021ja030071

Picone, J. M., Hedin, A. E., Drob, D. P., & Aikin, A. C. (2002). NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *Journal of Geophysical Research*, *107*(A12), 1468. https://doi.org/10.1029/2002JA009430

Rakipi, A., Kamo, B., Cakaj, S., & Lala, A. (2015). Standard positioning performance evaluation of a single-frequency GPS receiver implementing ionospheric and tropospheric error corrections. *International Journal of Advanced Computer Science and Applications*, *6*(3). https://doi.org/10.14569/ijacsa.2015.060304

Reinisch, B. W. (1995). *The Digisonde network and databasing*. World Data Center A for Solar-Terrestrial Physical, Report UAG-104 (pp. 8–15). Ionosonde Networks and Stations.

Reinisch, B. W., & Galkin, I. A. (2011). Global ionospheric radio observatory (GIRO). *Earth Planets and Space*, *63*(4), 377–381. https://doi.org/10.5047/eps.2011.03.001

Reinisch, B. W., & Galkin, I. A. (2023). FastChar [Dataset]. Retrieved from https://giro.uml.edu/didbase/scaled.php

Rideout, B. (2023a). Cedar [Dataset]. Retrieved from http://cedar.openmadrigal.org/single?isGlobal=True&categories=17&instruments=8000

Rideout, B. (2023b). Millstone [Dataset]. Retrieved from http://millstonehill.haystack.mit.edu/single?isGlobal=on&categories=1&instruments=30&years=2012&months=1&days=15

Scherliess, L., Schunk, R. W., Sojka, J. J., Thompson, D. C., & Zhu, L. (2006). Utah State University global assimilation of ionospheric measurements Gauss-Markov Kalman filter model of the ionosphere: Model description and validation. *Journal of Geophysical Research*, *111*(A11), A11315. https://doi.org/10.1029/2006JA011712

Schreiner, W. S., Sokolovskiy, S. V., Rocken, C., & Hunt, D. C. (1999). Analysis and validation of GPS/MET radio occultation data in the ionosphere. *Radio Science*, *34*(4), 949–966. https://doi.org/10.1029/1999rs900034

Shim, J. S., Kuznetsova, M., Rastätter, L., Bilitza, D., Butala, M., Codrescu, M., et al. (2012). CEDAR Electrodynamics Thermosphere Ionosphere (ETI) Challenge for systematic assessment of ionosphere/thermosphere models: Electron density, neutral density, NmF2, and hmF2 using space based observations. *Space Weather*, *10*, S10004. https://doi.org/10.1029/2012SW000851

Stankov, S. M., Verhulst, T. G. W., & Sapundjiev, D. (2023). Automatic ionospheric weather monitoring with DPS-4D ionosonde and ARTIST-5 autoscaler: System performance at a mid-latitude observatory. *Radio Science*, *58*(2), e2022RS007628. https://doi.org/10.1029/2022RS007628

Themens, D., Reid, B., & Elvidge, S. (2022). ARTIST ionogram autoscaling confidence scores: Best practices. *URSI Radio Sci. Lett*, *4*, 1–5.

Vierinen, J., Coster, A. J., Rideout, W. C., Erickson, P. J., & Norberg, J. (2016). Statistical framework for estimating GNSS bias. *Atmospheric Measurement Techniques*, *9*(3), 1303–1312. https://doi.org/10.5194/amt-9-1303-2016

Vines, S. K. (2023). Active magnetosphere and planetary electrodynamics response experiment [Dataset]. Retrieved from https://ampere.jhuapl.edu/download/?page=dataTab

Wang, H., Akmaev, R. A., Fang, T. W., Fuller-Rowell, T. J., Wu, F., Maruyama, N., & Iredell, M. D. (2014). First forecast of a sudden stratospheric warming with a coupled whole-atmosphere/ionosphere model IDEA. *Journal of Geophysical Research: Space Physics*, *119*(3), 2079–2089. https://doi.org/10.1002/2013ja019481

Welling, D. T., Jordanova, V. K., Glocer, A., Toth, G., Liemohn, M. W., & Weimer, D. R. (2015). The two-way relationship between ionospheric outflow and the ring current. *Journal of Geophysical Research: Space Physics*, *120*(6), 4338–4353. https://doi.org/10.1002/2015ja021231

Wübbena, G., Schmitz, M., & Bagge, A. (2005). PPP-RTK: Precise point positioning using state-space representation in RTK networks. In *Proceedings of ION GNSS 2005*.