ARTIST Ionogram Autoscaling Confidence Scores: Best Practices

David R. Themens, Ben Reid, and Sean Elvidge

Abstract – Ionogram data from the Global Ionospheric Radio Observatory (GIRO), processed with Automatic Real-Time Ionogram Scaler With True Height (ARTIST), is provided with confidence scores to guide users on the reliability of the data set. Here we use manually processed data to assess the extent to which these confidence scores have value and to provide users with guidelines on how best to apply these confidence scores when filtering GIRO data. It is found that the best confidence score for a given ionospheric parameter (foF1, foF2, or hmF2) is not necessarily the highest confidence score, and depends heavily on the parameter in question. In addition to this we demonstrate an 11.4% false negative rate in ARTIST's scaling of the F1 layer trace, missing nearly 40% of observed F1 traces.

1. Introduction

Ionosondes provide one of the longest-running and most reliable measurements of the ionosphere [1]. By scanning across a range of frequencies in the HF band and measuring the echo time of signals reflected from the ionosphere, they are capable of producing ionograms. These can be inverted into profiles of the ionosphere's electron density from the lower E-region to the peak of the F-region [2]. In order to determine electron density from an ionogram, however, one must first isolate ionospheric echoes from interference, separate the ordinary (O) and extraordinary (X) modes of propagation, and account for complex propagation effects due to nonisotropic ionospheric structuring, in a process referred to as "scaling." This is particularly true for high-latitude regions, where small-scale structures and high ionospheric drift speeds can make ionograms complex.

Before the advent of automatic scaling algorithms that conduct the trace isolation (scaling) process, scaling of ionograms was done manually by experienced ionosonde operators. The process of manual scaling has often been referred to as an art form in the field, requiring extensive experience and training. The practice was thereby trained at various workshops, with manuals and standards that were developed to ensure the consistency and reliability of the data set [3]. As time has progressed, however, the availability of automatic scaling software has provided a compelling alternative to the painstaking and time-consuming process of manual scaling. Today, few of the devoted manual-scaling experts of the past remain active in the field, and opportunities for training are rare. This lack of manual scalers has resulted in a substantial reliance on automatic scaling routines; the most popular is Automatic Real-Time Ionogram Scaler With True Height (ARTIST), provided with the Digisonde brand of ionosondes, due in part to the ease of use and availability of data from those systems through the Global Ionospheric Radio Observatory (GIRO; https://giro.uml.edu/) [4, 5].

To provide users with a measure of quality assurance, confidence scores were introduced and are distributed with all GIRO data. These scores range from 0 to 100 and are calculated by subtracting a penalty from the perfect score (100) if certain complicating ionospheric features are present or error conditions are experienced [6].

It should be noted that while the penalty criteria are provided as a list in SAO.XML versions of ARTIST output, they are not provided in any other form of data distributed from GIRO; instead, only a total score is provided. This lack of specification of the penalty criteria with the scores in the GIRO Quick Characteristics or SAOExplorer Characteristics output files leads to a degree of ambiguity in the nature of the scores. As most users simply apply a threshold on these scores to remove potentially problematic data (say, removing all data with confidence score below 70), the nuance of these scores is overlooked and very different error behaviors can be lumped together with the same quality score, despite having varying implications on different elements of the scaled profiles.

In this study, we use data from approximately 35 000 manually scaled Digisonde ionograms to assess the behavior and utility of the ARTIST confidence scores. We will further use these results to provide the community with best practices and guidance on their application.

2. Data and Scope

We have here limited this study only to data by the principal author on a touch-screen laptop within the past two years, to ensure a measure of scaling consistency. This data set is thereby limited to only 34 968 ionograms. The principal author has extensive scaling experience, with both Digisonde and Canadian Advanced Digital Ionosonde (CADI) ionograms, having scaled a few million ionograms and having been the

Manuscript received 15 January 2022.

David R. Themens is a Lecturer of Space Systems and Space Weather in the Space Environment and Radio Engineering Group (SERENE) at the University of Birmingham, UK. He is also an Adjunct Professor in the Department of Physics at the University of New Brunswick, Canada. email: d.r.themens@bham.ac.uk

Sean Elvidge is an Associate Professor of Space Environment in the Space Environment and Radio Engineering Group (SERENE) at the University of Birmingham, UK. email: s.elvidge@bham.ac.uk

Ben Reid is a PhD student in the Department of Physics at the University of New Brunswick, Canada. email: ben.reid@unb.ca

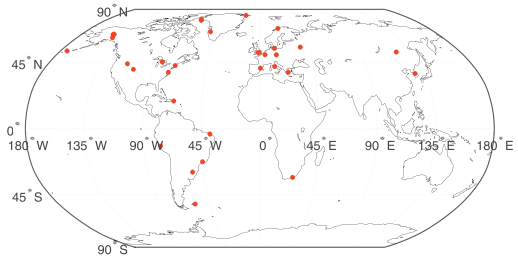

Figure 1. Distribution of GIRO Digisondes used in this study.

main processor of CADI data from the Canadian High Arctic Ionospheric Network (CHAIN) for over a decade [7]. Ionograms were scaled according to the *URSI Handbook on Ionogram Processing* [8] and supplemented with the recommendations of the *URSI High Latitude Supplement* [9] for the included high-latitude locations. A map of the distribution of stations used in this study is provided in Figure 1.

It should be noted that the data set used here is not evenly distributed spatially, and the data do not necessarily reflect the same time periods across all stations; in fact, much of the data set here is heavily concentrated in the three periods 12–20 March 2013, 14–17 July 2017, and 19–23 September 2017, where the majority of the stations in Figure 1 were scaled if data was available. Other periods generally only have one to three stations scaled coincidently. Although the data set is not ideally distributed, the intention of this study is to examine the confidence scores of ARTIST rather than to draw explicit conclusions regarding the accuracy of ARTIST as a whole, which would likely require a more extensive and evenly distributed data set.

We have here limited the scope of our study to foF1, foF2, and hmF2. We examine foF1 because it represents one of the most challenging parts of the ionogram to scale correctly using an automated method, as many features can be easily mistaken for an F1 trace if care is not taken and standard scaling protocols outlined in [8] are not adequately followed. We examine foF2 for its sheer popularity and importance in a number of applications. Finally, we examine hmF2 for similar reasons as foF2. It should be noted that hmF2 is very rarely assessed in existing studies of autoscaler performance [10, 11], making its accuracy somewhat of an unknown. The lack of such comparisons is often because the number of assumptions made in the full profile inversion process makes an estimate of error not entirely reflective of the actual performance. A number of conflating factors, such as assumptions regarding missing low-frequency portions of the trace or regarding the E–F valley region, can make a complete estimate of hmF2 performance challenging, depending heavily on the location of the instrument, inversion software, and even geophysical conditions. Here, as we are using the GIRO SAOExplorer software, which uses the same inversion method and assumptions as ARTIST, errors reflected in our analysis are indicative of just autoscaling performance and not the full hmF2 error, which may be larger than the values reported here due to the aforementioned assumptions and inversion errors.

3. Results and Discussion

This study seeks to provide answers to the simple question: Are the ARTIST confidence scores useful, and if so, how should they be used? To address this question we begin by first examining the performance efficiency (PE) of ARTIST for all confidence scores. PE is a skill score that measures the mean squared error of a model against the use of an average of the observations. If the PE is unity, the model is a perfect representation of the data, and if the PE is zero or negative, then the model performs worse than a simple average of the observations. The expression for PE is

$$\mathrm{PE} = 1 - \frac{\sum (M_i - O_i)^2}{\sum (O_i - \bar{O})^2} \tag{1}$$

where M_i is the model value, O_i is the observation, and $\bar{O}$ is the mean of the observations [12]. One can also replace the mean of the observations in the denominator with another reference model to assess the performance

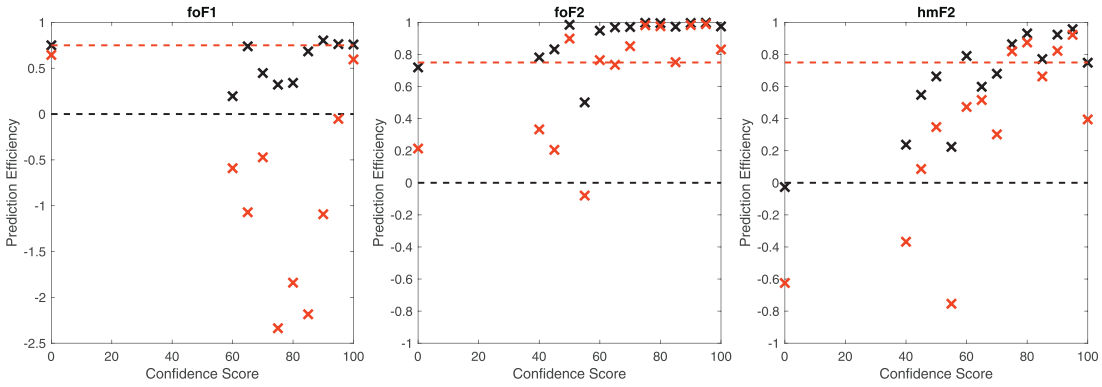

Figure 2. Prediction efficiency with respect to the mean (black) and the IRI (red) against confidence scores for foF1 (left), foF2 (middle), and hmF2 (right). Dashed lines are plotted at the 75% (red) and 0% (black) thresholds. Note the change in y-axis range for the foF1 subplot.

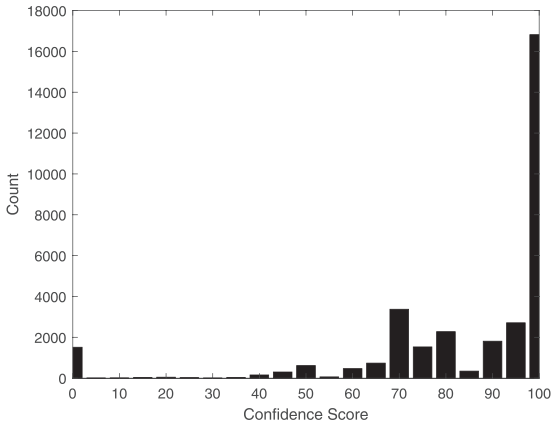

Figure 3. Occurrence rate of individual confidence scores.

with respect to that reference model instead of the mean, where negatives would then imply that the model is generally worse than the reference model.

To assess the value of the confidence scores and their relative performance, we determine the PE both with respect to the mean and with respect to the International Reference Ionosphere (IRI) [13]. Default IRI2016 options were used. In Figure 2, we present the Pes for foF1, foF2, and hmF2, doing a separate PE calculation for each possible confidence score. It should be noted that PE values are reported only if there were at least 100 measurements at the corresponding confidence score.

We note from Figure 2 that the PE generally decreases with decreasing confidence score, which suggests that the confidence scores do hold value in allowing users to assess the reliability of these characteristics; however, the range of confidence scores that provide reliable measurements is different for each ionospheric characteristic. First considering results with respect to the mean, for foF1 we see that confidence scores are generally lower than those for foF2, and the confidence scores corresponding to reliable foF1 are far more limited. Using a threshold of 75% (0.75) for PE, scores of 0, 90, 95, and 100 correspond to the most reliable foF1 estimates. For foF2, the range of reliable confidence scores expands significantly to any score of 40 or greater except the score of 55, which performs the worst of all confidence scores. For hmF2, scores of 60, 75, 80, 85, 90, and 95 correspond to PEs greater than 75% when calculated with respect to the mean. It is interesting to note here that the confidence score of 100 falls below the 75% PE level for hmF2 while generally performing well for the frequency parameters.

Figure 2 also shows the PEs calculated with respect to the IRI. For foF1, only confidence scores of 0 or 100 provide any skill with respect to the IRI; all other scores perform worse than the climatological model. For foF2, we see similar results to those with respect to the mean; however, we now note that all scores below 50 and the single score of 55 provide little skill over climatology and are largely unreliable. Finally, for

Table 1. Contingency table of manual scaling vs. autoscaling occurrence of the F1 layer

		Manual	
	F1 Layer	Present	Absent
Automatic	Present	6184	505
	Absent	3961	24318

hmF2, confidence scores of 75, 80, 90, and 95 provide excellent skill with respect to climatology, but all other scores provide limited skill improvement or, in the case of scores 0, 40, and 55, actively produce worse estimates of hmF2 than the climatological IRI. Circumstances like this may account for some of the reduced performance of the Real Time IRI (IRTAM), which assimilates these GIRO data, with respect to the climatological IRI in terms of hmF2 found in [14].

It is important to also contextualize these results with the relative frequency of each confidence score. To assess this, the frequency of observed confidence scores for the tested data set is provided in Figure 3.

We note here that the 100 confidence score forms nearly half of the scaled data set; therefore, for hmF2—which sees ARTIST performing at only 40% better than climatology for the 100 confidence score—discarding all ARTIST measurements below the 75% PE (improvement) level would result in discarding 74.7% of the scaled record. For many applications, having more, poor-quality data may be more valuable than discarding such a large portion of the available data set; however, if that is the case for that application, scores of 50 and 60 should be regarded as providing similar value and should thereby be retained for that application. Similarly, for foF1, a confidence score of 0 actually provides more value than one of 100, and therefore a desire to retain measurements of foF1 with a confidence score of 100 should be paired with retention of values at 0 as well.

While the PE of each confidence score is a valuable measure of performance when values are measured, it lacks another form of error, namely the scaling of ionograms or ionogram features when those features are not present or when ionogram simply cannot be scaled. As PE can only be calculated when there is both a manually scaled and autoscaled value available, it does not reflect errors in event occurrence, which are particularly important in the context of foF1. Contingency tables for scaling the F1 and F2 layers are provided in Tables 1 and 2.

From these contingency tables we note that ARTIST has an 11.3% false negative rate, where it does not scale the F1 layer when one is present, and a

Table 2. Contingency table of manual scaling vs. autoscaling occurrence of the F2 layer

		Manual	
	F2 Layer	Present	Absent
Automatic	Present	30406	1648
	Absent	999	1915

respectable 1.4% false positive rate, where it scales a feature as the F1 layer when one is not present. This implies that ARTIST is rather conservative in its identification of the F1 layer, which is understandable given the number of alternative features that may be mistaken for an F1 trace; however, the conservative scaling of the F1 layer has here resulted in ARTIST missing 39.0% of observed F1 layer traces. The most common error in scaling the F1 layer was ARTIST scaling the F1 trace as the F2 trace by accident, most commonly at high latitudes, where there is a strong F1–F2 cusp, and during storms where G-conditions often see the F2 layer recede behind the F1 layer.

For the F2 layer the contingency table shows only modest false negative (2.9%) and false positive (4.7%) rates; however, this is somewhat expected, as outside of severe events or extremely low-density periods, an F2 layer is almost always present (89.8% of the time). False positives were often made when blanketing sporadic-E features were mistaken as F2 traces or when ionograms were of severely poor quality and scaling could not be undertaken at an URSI Qualifying Letter level of U or better, most often the case at high latitudes. False negatives were less common and did not appear to have any particularly consistent behavior.

4. Conclusions

We have found that ARTIST confidence scores do provide a measure of value in characterizing the reliability of ARTIST-scaled characteristics; however, this value is strongly dependent on the characteristic of interest. It is shown here that for foF2, confidence scores as low as 50 (excluding 55) all correspond to reliable autoscaling performance for that parameter. For hmF2, 75, 80, 90, and 95 all provide significant value with respect to using climatological values; however, confidence scores of 0, 40, and 55 were actively worse than using the IRI, and values other than these provided only limited improvement. For foF1, virtually all confidence scores were worse than the IRI, with only scores of 0 and 100 providing any skill over the IRI, and both below 60% improvement.

It is further shown that ARTIST struggles somewhat in identifying the presence of the F1 layer, demonstrating a false negative rate of 11.3% and a false positive rate of 1.4%. This significant false negative rate amounts to 39% of all observed F1 layer occurrences and makes it challenging to assess the reliability of foF1 using conventional root-mean-square statistics, which only account for periods where both measured and autoscaled values are available.

Ultimately, we conclude that filtering ARTIST data available through GIRO is not a trivial matter and depends heavily on the parameter of interest. Although we here limit our analysis to only information that would be available from GIRO's Quick Characteristics page (https://giro.uml.edu/didbase/scaled.php), which is the dominant source of autoscaled ionosonde characteristics used for scientific purposes, future work will examine the overall performance of ARTIST autoscaling, compare confidence scores from ARTIST with those of Qualscan [15], and break down confidence scores by penalty criteria for a more in-depth examination of autoscaling error behavior and confidence. The intention here is simply to provide users with guidance on how best to capitalize on the value of the existing ARTIST confidence scores and avoid misusing them.

5. Acknowledgments

We thank the many ionosonde operators who contribute to GIRO for the use of their data for this study (http://spase.info/SMWG/Observatory/GIRO).

6. References

1. E. Araujo-Pradere, E. C. Weatherhead, P. B. Dandenault, D. Bilitza, P. Wilkinson, et al., "Critical Issues in Ionospheric Data Quality and Implications for Scientific Studies," *Radio Science*, **54**, 5, May 2019, pp. 440-454, doi: 10.1029/2018RS006686.
2. E. Titheridge, "The Real Height Analysis of Ionograms: A Generalized Formulation," *Radio Science*, **23**, 5, September-October 1988, pp. 831-849, doi:10.1029/RS023i005p00831.
3. P. Wilkinson, "IPS Scaling Conventions," https://www.sws.bom.gov.au/IPSHosted/INAG/scaling/ips_scaling_conventions03_1996.pdf (Accessed 26 December, 2021).
4. A. Galkin, G. M. Khmyrov, A. V. Kozlov, B. W. Reinisch, X. Huang, et al., "The ARTIST 5," *AIP Conference Proceedings*, **974**, 1, February 2008, pp. 150-159, doi: 10.1063/1.2885024.
5. W. Reinisch and I. A. Galkin, "Global Ionospheric Radio Observatory (GIRO)," *Earth, Planets and Space*, **63**, April 2011, pp. 377-381, doi: 10.5047/eps.2011.03.001.
6. A. Galkin, B. W. Reinisch, X. Huang, and G. M. Khmyrov, "Confidence Score of ARTIST-5 Ionogram Autoscaling," INAG Technical Memorandum, https://www.ursi.org/files/CommissionWebsites/INAG/web-73/confidence_score.pdf (Accessed 26 December, 2021).
7. T. Jayachandran, R. B. Langley, J. W. MacDougall, S. C. Mushini, D. Pokhotelov, et al., "Canadian High Arctic Ionospheric Network (CHAIN)," *Radio Science*, **44**, 1, February 2009, p. RS0A03, doi: 10.1029/2008RS004046.
8. R. Piggott and K. Rawer (eds.), *URSI Handbook of*

Ionogram Interpretation and Reduction, 2nd Ed., Revision of Chapters 1-4, Report UAG-23A, Boulder, CO, World Data Center A for Solar Terrestrial Physics, 1978, https://www.sws.bom.gov.au/IPSHosted/INAG/uag_23a/UAG_23A_indexed.pdf.

9. R. Piggott, *High-Latitude Supplement to URSI Handbook of Ionogram Interpretation and Reduction*, Report UAG-50, Boulder, CO, World Data Center A for Solar Terrestrial Physics, 1975, https://www.sws.bom.gov.au/IPSHosted/INAG/uag_50/uag_50.html.
10. M. Stankov, J.-C. Jodogne, I. Kutiev, K. Stegen, and R. Warnant, "Evaluation of Automatic Ionogram Scaling for Use in Real-Time Ionospheric Density Profile Specification: Dourbes DGS-256/ARTIST-4 Performance," *Annals of Geophysics*, **55**, 2, June 2012, pp. 283-291, doi: 10.4401/ag-4976.
11. C. Scotto and D. Sabbagh, "The Accuracy of Real-Time h_mF2 Estimation From Ionosondes," *Remote Sensing*, **12**, 17, August 2020, p. 2671, doi: 10.3390/rs12172671.
12. H. Murphy, "Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient," *Monthly Weather Review*, **116**, 12, December 1988, pp. 2417-2424, doi: 10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.
13. D. Bilitza, D. Altadill, V. Truhlik, V. Shubin, I. Galkin, et al., "International Reference Ionosphere 2016: From Ionospheric Climate to Real-Time Weather Predictions," *Space Weather*, **15**, 2, 2017, pp. 418-429, doi: 10.1002/2016SW001593.
14. A. Pignalberi, M. Pietrella, and M. Pezzopane, "Towards a Real-Time Description of the Ionosphere: A Comparison Between International Reference Ionosphere (IRI) and IRI Real-Time Assimilative Mapping (IRTAM) Models," *Atmosphere*, **12**, 8, August 2021, p. 1003, doi: 10.3390/atmos12081003.
15. F. McNamara, "Quality Figures and Error Bars for Autoscaled Digisonde Vertical Incidence Ionograms," *Radio Science*, **41**, 4, August 2006, p. RS4011, doi: 10.1029/2005RS003440.