Response to Reviewer 1's comments

We thank the reviewer for their comments and careful reading of the manuscript. In the following, the reviewer's comments are in black and responses are colored in blue. Where context is needed, paragraphs starting with ">>" indicate the reviewer's previous comments; paragraphs starting with ">" indicate the authors' previous responses.

Section 3.1 was improved significantly and is now better readable also for a non-expert of the field. However, Section 3.2, which was mentioned in the detailed comments, should be improved similarily. Please give references to, or explain terms like "batch size", "epoch", "Adam optimizer", "learning rate", LLRD, and BERT. Many of these are basic AI terminology, but references would be very useful for non-experts in the field.

Response: Section 3.2 has been expanded to:

'The model is trained using about 2 million synthetic samples and validated on 10,000 synthetic validation samples. The training uses mini-batches with a batch size of 512. A mini-batch is a small chunk of the training data that is used to do one model update. Larger batches improve computational efficiency but increase memory usage. Smaller batch size uses less memory and sometimes leads to better generalizations, at the cost of training speed. In machine learning, generalization means the model's performance on new, unseen data. We train for 30 epochs, where one epoch denotes one full pass through the entire training set. Model parameters are optimized with the Adam optimizer (Kingma and Ba, 2015), an adaptive first-order stochastic optimization method widely used for deep networks, and mean squared error (MSE) as the cost function. We use a cosine learning rate schedule (cosine annealing) with a linear warmup, i.e., the learning rate increases linearly from 10^{-7} to 10^{-4} during the first 1000 iterations, then decays following a cosine curve back to 10^{-7} by the final epoch. Learning rate is the step size the optimizer uses when it updates the model parameters. Cosine schedules (Ilya and Hutter, 2016) are commonly used to improve optimization stability and final convergence.

Layer-wise Learning Rate Decay (LLRD, Jeremy and Ruder, 2018) is a fine-tuning strategy. It assigns smaller learning rates to lower (earlier) layers and larger learning rates to higher (later) layers, reflecting the idea that early layers often learn more general representations while later layers are more task-specific. LLRD is widely used in natural language processing. In our application, we find LLRD is important for stable training, particularly for deeper architectures. Without LLRD, increasing the number of transformer blocks often degrades performance (i.e., deeper models fail to benefit from

scaling). While our model can operate with as few as one transformer block, we tested depths up to 100 blocks and observed consistent performance gains across this range. We therefore choose a 31-block architecture as a practical trade-off among hardware compatibility, inference speed, and model performance, and use LLRD with a decay rate of 0.9 per block to stabilize training and support scaling.'

We have also included additional references.

Thank you for the clarification. The Doppler shift grid resolution of 0.1 m/s also seems sufficient, but how wide is this grid? Also, are the biases in Figure 3 (a) calculated as mean values over all fit results at each altitude? If the grid is not wide enough, the posterior distribution may be non-zero at its edges. This will cause bias in mean values of the fitted velocities and their variances when the true velocity is non-zero, because the distribution is truncated at different distances from its true mean value at negative and positive sides. Maximum value of the posterior distribution should still provide an unbiased estimate.

I am still surprised to see that bias and standard deviation of the LSF depends on velocity in Figure 3 (a) and (b). Such behaviour might perhaps be expected for the moment method, but I cannot see how zero-mean noise added to the spectra could bias the LSF. The altitude variations make a qualitative match with the idea that they may be caused by truncation of the posterior distribution due to insuffucient width of the velocity grid. The authors should investigate the peak values of the posterior distribution or test the analysis with a wider velocity grid to see if this affects the results.

As an example, based on Figure 3, the true velocities reach almost 60 m/s and standard deviation of the fit reaches 15 m/s at some altitudes. To reach 4-sigma limit on both sides of the distribution at all altitudes, which is safe for sure, the velocity grid should span the range [-120 120] m/s. If the range is reduced to [-90 90] m/s (2-sigma), almost 3 percent of the posterior distribution may be cut off, and the mean values will be biased accordingly.

Response:

1. The velocity grid is [-128, 128] m/s. This range is sufficient for all altitudes and noise levels tested, so the posterior is not truncated by the grid. Therefore, truncation by the velocity grid does not occur in our simulations and does not explain the observed bias.

2. The LSF dependence on noise is expected. On average, LSF tends to underestimate the speed magnitude, and the underestimation becomes stronger as noise increases. A quick thought experiment is, in the extreme noise-limit case where the signal barely exists, the cost function carries little information about the true velocity, and the fitted velocity becomes dominated by noise and the fitting window. The resulting estimates are approximately symmetric over the fitting range and the mean velocity tends to zero. Therefore, when the true velocity is nonzero, the average estimate is pulled toward zero, which is an underestimation in magnitude. When the noise does not dominate, the bias is smaller but still expected, because the cost surface is still broadened by noise and the estimate is partially biased toward zero. This behavior is consistent with the trend in Fig. 3.

>> 5. The Arecibo radar collapsed a few years ago, but there are several other incoherent scatter radars in the world. Re-analysis of the archived Arecibo data is indeed valuable, but the authors could also comment if their technique might be usable for data from other radars that have considerably lower SNR and operate in completely different geophysical environments. In particular, other radars may observe much larger velocities and the users are typically interested also in electron densities and electron and ion temperatures, not just plasma velocities. At very end of the conclusions the authors claim, without any justification, that the model can be applied more broadly, but the very different noise levels and very much larger line-of-sight velocities observed with many other ISRs are not discussed at all.

> Response: Although the examples in this study are designed specifically for Arecibo ISR, the model itself is trained entirely on synthetic ISR spectra, for which radar parameters, Doppler velocity range, and SNR are manually controlled. Therefore, the training set can be easily adapted to a different SNR and Doppler velocity range.

> The relevant text has been revised to 'As the training data are generated using physics-based ISR simulations, SNR and Doppler velocity range can be explicitly controlled during data generation. Therefore, the proposed framework is not inherently limited to the Arecibo ISR and can be adapted to other instruments by retraining the model using instrument-specific parameters and configurations.'

The training set could be generated for any radar system and conditions, but this does not guarantee good performance in data analysis. Very much larger velocities, lower SNR, and steep gradients in plasma parameter profiles are detected by high-latitude radars. I agree with the reply above, but I do not find justification for the much stronger statement "The AI approach applies to all situations where the spectrum can be parameterized." in the abstract and "the approach applies to any situation where the observed spectra can be parameterized." in Conclusions. Please give some evidence that the method actually works in the very different conditions where many other ISRs operate, or change these sentences.

Response: There are two aspects to the issue. The first one is whether the AI approach can be applied to “all situations where the spectrum can be parameterized”, including non-ISR situations. The answer here is “yes”. One can always train the AI model using the theoretical spectra and make inferences. This is what we intended in the abstract. The second issue is whether AI can do better than LSF (or other methods) “in all situations”. The shorter answer is “we cannot prove this”. In the revision, we extend the discussion on the likelihood that AI outperforms LSF for other ISR measurements and other types of spectra in general. The conclusion is that AI will likely outperform LSF in the majority of situations. The reasons are given below:

Although our focus is on the Arecibo ISR at 430 MHz, the approach applies to other ISRs and non-ISR situations as long as the spectra can be parameterized. The AI model used here will likely outperform the LSF method in most of the other scenarios as well. In training the AI model for ISR applications, the only radar parameter that comes into play is the transmitting frequency unless the radar is pointed exactly perpendicular to the geomagnetic field line. Other radar parameters affect SNR, which can be adjusted by changing the height and/or time resolution. In any event, we have tested a large range of SNR, covering practically all the situations in which a useful Doppler can be obtained. Incoherent scatter fitting methods are generally known to be easily adaptable to different radars. There is no reason to believe that the advantages of the AI model discussed here do not apply to other ISRs.

Incoherent power spectra shapes are complex and diverse in the altitude range discussed here, including Gaussian-like and superpositions of Gaussian-like functions. If the model works for the ISR spectra, it will likely work for simpler non-ISR spectra as well, since those spectra are similar to subsets of the ISR spectra. Further, in dealing with sharp vertical changes in any of the state variables, such as the electron density in sporadic E’s, one can either reduce the number of heights in the awareness model or use a combination of the height-aware model and height-unaware model. Other than needing to train a different model, the AI model does not have any greater limitations than the LSF method in achieving different height resolutions.

We have added this discussion to the revised manuscript in the Discussion and Conclusion section.

The term "curve fitting method" is still used in several places. This should be replaced with "LSF method" everywhere.

Responses: Done. We searched the document and replaced all “curve fitting method” to LSF.

>> Equation (1): Shape of this profile seems to affect the final results, because the context-aware AI model learns this profile shape. Is there some physical justification for the selected function?

> Response: Equation (1) constrains the maximum vertical variation of plasma parameters over the 1.5 km altitude range, with hyperparameters selected empirically based on variability observed in real ISR measurements.

This is a very critical point for the analysis and should be explained in detail in the manuscript. Please add a detailed explanation about how Equation (1) was selected in the text. This selection may also restrict use of the model. For example, does it work in presence of narrow sporadic E layers, which produce very steep gradients in Ne profiles? Also, regarding applicability to other radar systems, how would steep gradients in Vi profiles affect the results?

Response:

We have expanded the manuscript to describe how Eq. (1) was selected and what variability it allows. Eq. (1) was introduced as a practical constraint on vertical variability within the 1.5 km window spanned by the five-height input. The maximum variation in the training data is up to 10% change over 1.5 km, which is a steep gradient for most conditions.

The following has been added to the end of the discussion for Eq. (1)

“Including a synthetic gradient makes the training profiles more realistic, but its impact on performance is limited unless N becomes large. For the relatively small N (=5, 1.5 km height resolution) used for the current work, the model is largely information-limited, so the structure introduced in Eq. (1) does not substantially change model performance. At the same time, over-representing extreme or non-representative gradient cases in the synthetic training set can shift the training distribution away from the target data and bias the model toward rare structures, which can degrade performance on typical observations. For this reason, the training set is designed to reflect the empirical distribution of gradients seen in real data.”

>> Lines 124-125: "In transformer architectures such as BERT or ViT (Devlin et al. 2019; Dosovitskiy et al. 2020)." This sentence seems to be completely detached from the surrounding text.

> Response: This was a typo in the original manuscript. It was supposed to be a comma rather than a period after the sentence.

The manuscript has not been changed accordingly.

Response: The typo is now fixed. A comma is placed after the sentence.

>> Line 199: "...context-unaware model is trained on standalone 101-point spectra with artificial noise..." Is this noise somehow different from the noise added to the 5x101 input of the context- aware model?
> Response: No. The same noise variance is applied independently at each height in both models. The context-aware model uses all five height-resolved spectra as separate tokens, while the context-unaware model averages the five heights.

The text was not modified and is still unclear. It gives the impression that the noise is somehow different for the context-unaware model. Please clarify the text.

Response: The text was revised to

“Both models use the same 31-block [CLS]-based architecture and the same underlying data (and noise), but are trained and tested with different input formats. The context-aware model receives the full 5 × 101 height-resolved spectra. The context-unaware model receives the same data after averaging the five heights, i.e., the 5 × 101 input is averaged to 1 × 101 and reshaped to 101 × 1.”

Thank you for clarifying this. However, it is not sufficient to put the explanation in the figure caption. It should be included in the main text at the point where LSF-ideal is first mentioned. Also, are the values rounded to the nearest grid point in the LSF, or are exact values used? If exact values are used, this might partially explain the differences between LSF-ideal and LSF-realistic.

Response: The Doppler values are rounded to the nearest velocity grid point in the LSF for both LSF-ideal and LSF-realistic due to the finite velocity grid used in the least-squares fitting.

The following text was added at the first mention of LSF-ideal:

“The LSF-ideal method outperforms the 5ht-aware model in the low noise regime, where the input spectrum is nearly noise-free and the fitting problem is well-conditioned. In this case, the spectrum is effectively a clean copy of the known target, and the algorithm can retrieve the Doppler largely without error, except for a small quantization error due to the finite velocity grid resolution. LSF-ideal and LSF-realistic differ only in how the spectrum used for Doppler fitting is obtained. In the LSF-ideal case, the true noise-free spectrum shape is assumed to be known, and the Doppler velocity is retrieved by shifting this fixed template along the frequency axis and minimizing the least-squares error. In the LSF-realistic case, the spectrum shape is unknown and must first be estimated from noisy data.”

At the limit of zero signal the distribution should become flat, and it is centered around zero only because the search space is symmetric with respect to zero velocity. For non-zero signals the distribution should be centered at the true velocity. It seems possible that distribution of the LSF results is truncated due to insufficient width of the velocity grid in this case (see my comment above).

Response: Yes, ultimately, the bias is a result of the center frequency of the passband being at 0 velocity and the passband’s finite bandwidth, which is necessary to reduce the statistical error. However, the velocity distribution when both signal and noise are present is not centered at the true velocity. We can think of the velocity estimate as a weighted average of the no-noise and no-signal results, with weights determined by the SNR. In the no-noise situation, one gets the true velocity. With noise added, the estimate is between the no-noise result (true velocity) and the no-signal result (0 mean).

Lines 107-108: "...where domain knowledge shapes the training data..."
Is "domain knowledge" the set of possible altitude profiles in equation (1) in practice?

Response: The domain knowledge refers to the theoretical incoherent scatter spectrum model.

The relevant text is revised to

‘This approach is consistent with broader definitions of physics-informed machine learning, where domain knowledge from ISR theory is used to generate and constrain the training data, rather than being explicitly hard-coded into the network architecture or loss function.’

Lines 130-131: "...Doppler shift manifests as a subtle, globally coherent displacement in frequency space that is shared across altitudes..."

The Doppler shift changes with altitude, it cannot be described as "globally coherent".

Response: The sentence is changed to "...as Doppler shifts are manifested as altitudinally coherent displacements embedded in spectra whose widths and amplitudes vary with local plasma parameters."

Line 147: AS -> As

Response: corrected

Line 161: "commonly referred to as [CLS] token in AI literature"
Please give a reference.

Response: Done

Title of Section 3.2. "training" -> "Training" (or perhaps "Training the AI model" or something similar?)

Response: The section title has been changed to 'Training the AI model'.

Line 255: "...When η is above 30 (~101.5), it is largely independent of..."
Does 'it' refer to the velocity RMSE?

Response: Yes. The relevant sentence has been changed to:

"When η is above 30 (~$10^{1.5}$), the velocity RMSE is largely independent of the equivalent bandwidth."

Line 276: "for a representative condition at Arecibo"
Please give details of these conditions.

Response: The following is added after the relevant sentence.

"The altitude-dependent noise variance profile is sampled from an Arecibo Observatory measurement on 11 September 2014, increasing from $10^{-4.2}$ at 90 km to 10^{-3} at 200 km."

Lines 350-351: "Above 120 km, a larger number of heights can be used in..."
Does one need re-train the model for this?

Response: Yes. While transformers can take a varying number of input tokens, i.e., a varying number of heights, in our setup, we keep the token length fixed because it is

simpler and tends to yield more stable training. Therefore, changing the number of heights changes the input length and the model must be retrained.

Response to reviewer 2's comments

The manuscript presents an application of transformer-based models to Doppler velocity estimation from incoherent scatter radar spectra, and the revised version shows improvements in clarity and organization. However, despite these improvements, the work still suffers from fundamental scientific and methodological deficiencies that prevent acceptance. The major issues are structural and require substantial revision.

The major concerns are below.

Issue 1: Poor scientific framing and overstated generality

The manuscript claims broad applicability of the proposed method, stating that it can be applied to any radar system or to any situation where the observed spectra can be parameterized. These claims are not supported by the results presented. The study demonstrates performance only for a specific ISR configuration and a single real-data case. Adapting the method to another radar system would require instrument-specific synthetic data generation, retraining of the model, and thorough revalidation. These steps are non-trivial and constitute a substantial methodological effort. This dependency is not adequately acknowledged in the manuscript, and no clear validity domain, assumptions, or potential failure modes are defined.

As a result, the conclusions are overstated relative to the evidence provided. The scope of applicability must be explicitly limited, and speculative statements regarding generalization should either be removed or clearly supported with demonstrations and results.

Response: We stand by our claim that the AI approach to estimate the Doppler shift can be applied to any radar system or to any situation where the observed spectra can be parameterized. The AI approach has two steps: 1) training, which can be accomplished as long as one knows the theoretical spectrum is determined by a finite number of parameters regardless the instrument; 2) inference, once the model is trained, the AI model will yield a prediction based on the observed spectra. The training and inference require the same amount of information as LSF. While the statement is true, we acknowledge that it does not address whether "AI can do better", which is likely what the reviewer has in mind. We have added the following discussion in the revised manuscript.

Although our focus is on the Arecibo ISR at 430 MHz, the approach applies to other ISRs and non-ISR situations as long as the spectra can be parameterized. The AI model used here will likely outperform the LSF method in most of the other scenarios as well. In training the AI model for ISR applications, the only radar parameter that comes into play is the transmitting frequency

unless the radar is pointed exactly perpendicular to the geomagnetic field line. Other radar parameters affect SNR, which can be adjusted by changing the height and/or time resolution. In any event, we have tested a large range of SNR, covering practically all the situations in which a useful Doppler can be obtained. Incoherent scatter fitting methods are generally known to be easily adaptable to different radars. There is no reason to believe that the advantages of the AI model discussed here do not apply to other ISRs.

Incoherent power spectra shapes are complex and diverse in the altitude range discussed here, including Gaussian-like and superpositions of Gaussian-like functions. If the model works for the ISR spectra, it will likely work for simpler non-ISR spectra as well, since those spectra are similar to subsets of the ISR spectra. Further, in dealing with sharp vertical changes in any of the state variables, such as the electron density in sporadic E's, one can either reduce the number of heights in the awareness model or use a combination of the height-aware model and height-unaware model. Other than needing to train a different model, the AI model does not have any greater limitations than the LSF method in achieving different height resolutions.

Issue 2: Lack of physical validation and bias toward smoothness

The manuscript does not provide convincing evidence of a physically validated improvement in Doppler velocity estimation. The evaluation using real data relies primarily on smoothness- and coherence-based metrics, such as second-order differences in altitude and time, which inherently favor smooth solutions. These metrics demonstrate only that the AI method produces smoother profiles than the least-squares fitting (LSF) approach.

In several instances, this smoothing appears to suppress physically meaningful structures. For example, in Figure 4b, around 115 km altitude near 12:00, a clear velocity structure visible in the LSF results is entirely removed (smoothed) by the AI method. This raises concerns that genuine physical variability may be attenuated or lost.

Response: We have demonstrated that the AI model reduces the error by ~50%. When comparing two data sets, "smoothness" is synonymous with statistical error. This is a fundamental aspect of statistics. In our case, the increased apparent "smoothness" is an expected consequence of reduced statistical variance (i.e., lower random error), not evidence of artificial smoothing.

When we do not know the ground truth and the results are within physical limits, there is no way to tell whether a feature is physical or not. We have demonstrated in simulation and real data that the statistical error from the AI model is smaller than LSF. We have also shown in simulations that the two methods have comparable bias. These two results convincingly demonstrate that AI outperforms the LSF method.

At the same time, the training data are synthetically generated using constrained, smooth vertical profiles. This creates a closed methodological loop in which the model is trained to favor smooth outputs and is subsequently evaluated using metrics that explicitly reward smoothness. Under these conditions, improved performance according to the selected metrics does not necessarily indicate improved physical accuracy, but rather stronger implicit regularization or denoising.

No independent physical validation is presented to demonstrate that the AI-derived velocities are closer to the true plasma drift than those obtained with least-squares fitting. The manuscript does not sufficiently distinguish between noise suppression and physical correctness, yet this distinction is essential for supporting the scientific claims being made.

Response: We presented two AI models, one does not have height awareness and the other has height awareness. The height-unaware model processes each height independently. As demonstrated in the manuscript, the height-unaware model still outperforms the LSF method with the same information. We did not focus on the height-unaware model because its advantage over LSF is not as pronounced as the height-aware model and the five-height-aware model has a fine enough height resolution (1.5 km) for our application.

When we compare the height-aware model with the LSF method, we use the same amount of information. The “smoothness” assumptions we use in training the height-aware model are general enough for the majority of the cases. In the event that the assumption breaks down, such as in a sporadic E, one would need to reduce the number of awareness heights or just use the height-unaware model. One can combine the results from the height-unaware model with those from the height-aware model. Ultimately, this is a height resolution problem. If one needs fine resolution, one cannot average too many heights either in LSF or let the AI model link too many heights together. AI models do not have more limitations than the LSF method in this regard other than that more AI models need to be trained.

Issue 3: Unjustified model complexity and unclear practicality

The proposed model is exceptionally large relative to the input size, consisting of 31 transformer blocks and approximately 100 million parameters. The manuscript does not provide a convincing justification for why such a large architecture is required, nor does it include comparisons with simpler models that could plausibly achieve comparable performance.

Furthermore, the discussion of computational efficiency is incomplete. While inference speed is emphasized, the substantial cost associated with model training and the reliance on modern GPU hardware are not properly contextualized. Least-squares fitting does not require a training phase and incurs relatively modest computational cost. As a result, the comparison between the two approaches is not balanced, and the practical advantages of the proposed method remain unclear.

Response:

The model is about 100M parameters, which is modest by current standards and fits on any reasonable modern GPU, so no specialized hardware is needed. We also include a scaling experiment (see figure below) showing that performance benefits from increased number of transformer blocks, e.g., the 100M model (~30 blocks) outperforms the 30M model (~10 blocks) under the same architecture and training. In the revised manuscript, we have added explicit training time, hardware, and inference throughput numbers to contextualize the trade-off.

Training the model takes about one day. While this is a non-trivial amount of time, it is a one-time cost. This upfront cost is small compared to the repeated runtime cost of LSF, which can take days to weeks to process a few days of radar measurements, whereas the AI model can process the same data in a few hours.