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.

The manuscript has again improved significantly with the unclear terminology now explained with sufficient references. The authors have also given detailed response to my comments about the bias in the LSF estimates.

I agree with the author's thought experiment about the fitted velocities being symmetric with respect to zero velocity at the limit of very low SNR, but the "pulling toward zero" must be caused by truncation of the (possibly very wide) distribution of the fitted velocities. To my understanding, the bias should thus be avoided with any non-zero SNR if wide enough velocity grid is used. However, the authors have also confirmed that the velocity grid is wide enough to cover the velocity distribution in this specific test case. The only remaining factor I can think of is the frequency grid used for the spectra calculations. This will not affect comparison of the analysis techniques because the same data are used with all analysis techniques.

My only remaining comment is therefore that the authors should check if the spectra are sufficiently close to zero at edges of the frequency grid when the velocity is 120 m/s. If this is not the case, the consequences on the velocity fits should be discussed accordingly. If the authors can confirm that the spectra are very close to zero at edges of the frequency grid in all cases, I recommend the manuscript to be accepted in its present form.

Response:

The general phenomenon of "pulling toward zero" is due to our selection of the applicable Doppler frequency range being centered at zero. If we choose the applicable frequency range (i.e., truncation range) of [0, 100] Hz, the expectation of the estimated frequency is 50 Hz in the case of zero SNR. The truncation range affects the standard deviation of the estimated Doppler. In comparing two methods, one needs to ensure that the truncation range is large enough to cover the limiting cases and is the same for both methods. The frequency resolution plays a negligible role as long as it's fine enough to have a reasonable number of points in the truncation range (e.g., 100).

Specific to our study here, the Doppler of the limiting velocity (120 m/s) is a small fraction of the spectral width and thus near the center of the spectrum. In the CLP configuration used here, the 500 kHz full bandwidth corresponds to a Doppler aliasing limit of about ±87.8 km/s. With the interpolated frequency spacing of 244.3 Hz, one frequency point corresponds to about 85.8 m/s. The limiting velocity of 120 m/s has only about 1.4 frequency points. Our fitting window has 101 frequency points, corresponding to about 24.7 kHz, which is not only sufficient to accommodate the limiting Doppler frequency but also the entire incoherent scatter spectral width (typically between 5 and 15 kHz wide for the Arecibo radar). The takeaway is that frequency clipping is not a concern in our configuration.

Response to Reviewer 2's comments

Reviewer's comments are in black. Responses are in blue.

The revised manuscript shows improvements in clarity, organization, and technical description. The authors have expanded several sections and clarified aspects of the transformer architecture, training strategy, and simulation framework. However, despite these revisions, the manuscript still suffers from fundamental scientific and methodological deficiencies that prevent acceptance in its current form. The major concerns remain largely unresolved and require substantial revision.

The main issues are summarized below.

Issue 1: Poor scientific framing, overstated generality, and lack of clearly defined scope and limitations

The manuscript continues to make broad claims regarding the applicability and generalization capability of the proposed AI approach, repeatedly stating that the method applies to "all situations where the spectrum can be parameterized" and can be extended broadly to other radar systems and non-ISR applications. These claims remain unsupported by the evidence presented.

The issue is not whether one can mathematically train a neural network on arbitrary spectra. Of course one can. The relevant scientific question is whether:

* the model generalizes robustly outside the training distribution,
* the AI approach consistently outperforms traditional estimators under realistic operating conditions,
* and whether the assumptions embedded in the training data remain valid under substantially different radar configurations, SNR regimes, and geophysical environments.

These points require experimental demonstration and validation, not plausibility arguments alone.

The manuscript demonstrates results only for a specific ISR configuration, a specific spectral parameterization, and a single real-data experiment from Arecibo. No cross-radar validation, no out-of-distribution testing, and no demonstration under substantially different ISR conditions are presented.

Given the limited validation presented, a critical missing component of the manuscript is a rigorous discussion of:

* scope,
* validity domain,
* limitations,
* and potential failure cases.

Furthermore, the authors suggests that different model configurations are preferable under different geophysical conditions, including combinations of height-aware and height-unaware approaches. However, this immediately raises unresolved scientific questions:

* Under what conditions does each model succeed or fail?
* What spatial or vertical scales are preserved or suppressed by each configuration?
* How can the user determine when the assumptions of the height-aware model break down?

At present, the manuscript does not provide a rigorous framework for answering these questions. As a result, the methodology appears heuristic and condition-dependent rather than systematically validated. The scope of applicability must therefore be substantially narrowed, or alternatively, the claims must be supported through additional validation experiments and a rigorous discussion of limitations and failure regimes.

Response: In the revision, we have limited the scope to the ISR application and merely pointed out the potential for other applications.

The input parameters to train the AI model are the variables affecting the ISR spectrum: including Doppler velocity, electron density, ion temperature and electron temperature, and the signal-to-nose ratio. The training distribution covers known limiting cases. Parameters related to the radar hardware are reflected in the signal-to-noise ratio and equivalent bandwidth as seen in Figure 3, which covers practically the entire useful operating range for UHF ISRs. The example from Arecibo represents a realistic case.

Issue 2: Lack of physical validation and suppression of small-scale variability

This remains the central unresolved scientific issue in the manuscript.

The manuscript demonstrates that the proposed AI model produces smoother and more “coherent” velocity fields than the LSF method. However, it still does not convincingly demonstrate that the resulting Doppler velocities are physically more accurate.

The evaluation of real data relies primarily on smoothness- and coherence-based metrics, particularly second-order differences in altitude and time. These metrics inherently favor smooth solutions. At the same time, the training data are generated using constrained smooth vertical profiles with limited variability. As a result, the methodology creates a closed loop:

* the model is trained to favor smooth outputs,
* and is then evaluated using metrics that explicitly reward smoothness.

Under these conditions, improved performance according to the selected metrics does not necessarily imply improved physical accuracy. It may instead indicate stronger implicit regularization or denoising.

The manuscript interprets increased smoothness as evidence of reduced statistical error and improved estimation quality. However, in statistical estimation, lower variance alone does not imply improved physical correctness. A strongly regularized estimator can reduce variance while simultaneously suppressing genuine small-scale variability and localized structures. Distinguishing denoising from physical fidelity therefore requires independent validation beyond smoothness- and coherence-based metrics.

This concern is particularly important because the proposed context-aware model explicitly links neighboring altitude bins and is trained using vertically smooth synthetic profiles. Such a framework naturally favors vertically coherent structures and may reduce sensitivity to localized variability or sharp gradients. Although the authors argue that the height-unaware model also performs well, the manuscript does not quantitatively characterize what spatial or vertical scales are preserved,

attenuated, or suppressed by either configuration.

As a result, it remains unclear:

* which classes of geophysical structures are faithfully reproduced,
* which are smoothed or attenuated,
* and under what conditions each model configuration should be preferred.

This issue becomes particularly important because the manuscript itself suggests combining the height-aware and height-unaware models depending on the observational scenario. However, such a strategy implicitly requires prior knowledge of when each model succeeds or fails, yet no quantitative framework or scale-dependent validation is provided to make this determination.

This behavior is already visible in Figure 4, where localized structures visible in the LSF solution appear substantially weakened or absent in the AI results. For example, around 115 km near 12:00 LT, a clear velocity feature visible in the LSF solution is almost completely suppressed by the AI model. The manuscript does not demonstrate whether these structures are noise artifacts or genuine physical variability.

From a geophysical perspective, the suppression of localized variability is itself a critical concern, particularly in regions where sharp gradients, intermittent structures, or small-scale dynamical processes are expected. A smoother solution is not necessarily a more physically accurate solution.

Response: We thank the reviewer for writing an extended report. There are, however, misunderstandings about the model, the training, and the evaluation metrics. The model only links 5 heights in the training. The model does not assume anything more than Eq. (1) for the five heights within the “aware” range. The LSF result for comparison is obtained by averaging over five heights as well so that the two methods use identical input information. For example, results from height 1-5 and height number 5-10 are independent for both methods. All the comparisons are made using identical information with the same independent height and time resolutions. **It is not true that “the (AI) model is trained to favor smooth outputs”.** We have modified the description in section 2 to provide further clarification and reduce confusion.

There is no "suppression of local variability" in either AI outputs or LSF results beyond the 5-heights used for AI-awareness or LSF average. The suppression of variability within the five aware heights is a matter of height resolution, and it is the same for both methods. If the output needs to be at the best resolution (but at the expense of a larger error), one can use the AI context-unaware model and LSF fitting without any height average. As discussed in the manuscript, AI still has an advantage, but it is not as pronounced as in the cases with height awareness.

We do not know what the reviewer means by "physical validation" and "physically more accurate" or how they can be potentially done. We compare the two methods by using controlled data with known inputs. We show that the AI model offers smaller variance and has a comparable bias to the LSF method in Section 3. For physical data taken at Arecibo, we do not know the ground truth. In the absence of ground truth, we can only compare the variance/standard deviation. We show that the AI model has a smaller variance. With comparable biases, the method with a smaller variance is considered more "accurate". **The two sets of comparisons presented in the manuscript establish that AI is the more accurate method under known statistical analysis norms**.

In all the comparisons, we have ensured that LSF and AI use identical information, and the LSF and AI outputs have the same height and time resolutions. Standard deviations are computed using only independent samples, which is emphasized in the revision (even though this is not a necessary condition to establish that one method is better than another).

The discussion in the manuscript using a varying number of heights at different altitude ranges in the AI model is similar to using a different window size for running average in the LSF method. This is a matter of adapting the algorithm to achieve the most desirable results prioritized by the users. As far as the comparison between the AI model and the LSF method is concerned, having more aware heights will favor the AI model even more, as discussed in the manuscript.