General comments:

The article describes a study of GIC forecasting using the spectral power of magnetic field perturbations as the GIC proxy target for LSTM based models. Using the spectral power is useful, as compared to magnetic field perturbations that are often used, because it has the capability to incorporate wave-driven phenomena and takes into account that the lithosphere and power grid are more susceptible to certain frequencies. This has the potential to be a useful contribution to the literature pending more details and discussion as described below.

Major specific comment:

Paragraph around Line 205: The data for the training set comes from January 2010 through September 2021, and the test set data comes from October 2021 through September 2024. The training set covers the beginning of the ascending phase of Solar Cycle 24 through the beginning of the ascending phase for Solar Cycle 25. It is generally desirable for a training dataset to constitute a full solar cycle so that it represents a complete range of solar and geomagnetic behavior.

However, the test set in this study is formed from data from a later time period than the test set. Data from the test set only represents the ascending phase of a solar cycle. Furthermore, the test data are from Solar Cycle 25, which, even by 2024, saw a higher occurrence of CMEs than in most of the previous solar cycle (https://helioforecast.space/solarcycle).

Could the authors clarify the reasoning behind the year-based split between the training-validation set and the test set, as opposed to, for example, a split in which the test set is a random 10% of the data from January 2010 through September 2021, and the training-validation set is the other 90%?

Given the authors' year-based split, is the test set representative of the entire dataset, in a statistical sense? If it is not, evaluating the model on the test set may give a false impression of the model's performance. The authors should either demonstrate that the test set is representative of the entire dataset, or determine a method to make it representative, perhaps by sampling the training-validation set and the test set from the same time period, as described above.

In this particular case, the test set has more extreme storms than the training set, yet the models perform surprisingly well. For example, Fig 6 for the May 2024 storm has reasonable magnitudes compared to observations, which isn't often seen in ML models. Do you have any insight to this? This is where more discussion of the training/testing on more comparable datasets could be useful.

Minor specific comments:

Line 37: It would enhance the clarity of this point if the authors would state a range of speeds typical of solar wind (or of both quiescent solar wind and geoeffective solar wind).

Line 54: Do these forecasts' "perfectness" come from their actual quality, or rather from the idea that we should ignore their uncertainties for the sake of this study? Consider adding to this sentence something like: "*The latter are "perfect" in the sense that they are based on actual solar wind and IMF observations and CME arrival times at Earth, ignoring any uncertainties in those observations."*

Line 98: The authors mention that this distribution has a standard deviation of 5 to 6 hours, but what determines the exact value of the standard deviation for a given case?

Table 1 log(P0) is listed as an input, but it's also the target. Do you just mean the time history of this variable is an input or something else?

Line 160-The text says m is the number in validation set. Shouldn't this be the number in whichever set you're calculating, not validation in particular?

Eq 4 Is the sum over n? If h=w, why are both variables used? Please clarify.

Line 164: What are the dropout probabilities? Were they tuned via the Optuna search, as described on lines 141-149? If not, how were they selected? It is recommended that these be discussed with that previous discussion of hyperparameters.

Section 3.3 There is no problem with benchmarking the LSTM against the hybrid persistence-climatological model, but how does the hybrid model fare against persistence only, or against climate only? It is not out of the question that one of those models could outperform the hybrid model, let alone the LSTM. Also, in line 348 a "recurrence model" is mentioned, which sounds like a persistence model using a 24 hour horizon. It would be helpful to be consistent with language and also mention that model in this section rather than much later.

Table 3: If the authors wish to express the confusion matrix elements (TP, FP, TN, and FN) as the number occurrences of each, as is currently done in this table, the total number of measurements, positives, and negatives should be provided in the caption or the header. Additionally, the caption only mentions the last two elements so needs much more description added.

Line 355 This statement is unclear, and possibly inaccurate: *deep learning models generally perform better than conventional machine learning models (e.g., neural networks and nonlinear models), which are only capable of predicting one step at a time* The examples given in the parentheses seem to be describing the former phrase, not the latter that the parentheses immediately follow. The limitation on predicting one step at a time also seems questionable.

Paragraph around line 370. Based on the discussion, it seems like training a model with CME arrival time excluded would also be a useful test. The conclusions here are also not clear. However, once I read Line 404, the conclusions make much more sense. *In the absence of any information about IMF B_z, CME arrival time is apparently unimportant.*

Certainly, if you don't know whether reconnection is going to be initiated (i.e. IMF Bz) then it doesn't matter what time it arrives. I think this discussion could be expanded upon for clarity.

Line 394 This sentence isn't clear. *Thus, the models' imperfection in space weather forecasting for longer forecast horizons was still reliable in forecasting events.*

The results for the longer forecast horizons aren't that great so don't seem like they're providing reliable forecasts, even if they're comparable to previous studies.

Line 396 This sentence is also unclear: *indicating reliable intelligence in the model benchmark by the persistence model.* The benchmark is a hybrid, not just persistence and could just say "model benchmark." However, it also sounds like it's saying the benchmark model has the "reliable intelligence." Are you trying to say "reliable intelligence in comparison to the benchmark"?

Figure 8, etc. Are the metrics calculated over the entire test set or just for the case study storms?

This is not the first study using AI to show the importance of IMF Bz. See, for example, Lotz and Cilliers, DOI:10.1016/j.asr.2014.09.014 and Coughlan et al, DOI:10.1029/2025SW004391

Technical corrections:

Line 20-21This seems to have two nouns in the final phrase.

Line 68, 144 etc-The article needs to be reviewed for correct citation type

Line 72: I looked up "first-order backward differencing" to clarify what is being done. This is okay but the authors could consider just putting the equation in.

Figure 2 labels and caption for b and c are switched

Line 153 *relatively much higher* pick “relatively” or “much”

Line 309 The formatting of the variables doesn’t match previous here.

There are a number of places where explanation/discussion comes later than it should. For example, at Eq 7 it’s unclear what MCC stands for but then it appears in the text after Eq 9. Second, the reason for using 0.985 comes later than where it is introduced. Third, the comparison to Hu 2024 comes first and then stating what Hu did comes later.