We thank the reviewer for their thorough evaluation of and constructive feedback on our manuscript. We propose several changes below that we believe address the reviewer's commentary and improve the manuscript. These changes are summarized in this letter, along with specific responses to the reviewer's comments. Below the reviewer's comments our response is shown in **bold**. Proposed modifications and/or additions to the manuscript are shown in *italics*.

Reviewer #1 Evaluation

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.

- **We appreciate the reviewer's overall positivity toward our approach, and agree that the points the reviewer has raised need revision/clarification.**

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).

- **We agree with the reviewer that it would be desirable to be able to include more data for training, validation, and testing. In fact, we originally envisioned covering the entire period from 2000 to 2024. Rørvik magnetometer data records prior to 2010 include chunks of missing readings and bad data that significantly complicates preparation of the data set, and we ultimately decided that there were more drawbacks than benefits. We therefore decided to work with data from 2010 to 2024, which contains virtually no instances of missing or bad data.**

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%?

- **This is an excellent question that deserves attention in the manuscript. The reason that we chose to split the data chronologically (instead of at random) is that there is a decent body of literature indicating that LSTM models trained on data sets that are split randomly suffer from data leakage, whereby information about the future is leaked into past measurements. A**

reasonably straightforward description of this problem is given by Sujeeth Kumaravel on a Medium webpage dedicated to this question: https://medium.com/@sujeeth.selvam/asdsadsad-3f690ca13d07. This is also discussed by IBM here: https://www.ibm.com/think/topics/data-leakage-machine-learning.

In the language of time series analysis, this problem can perhaps more intuitively be thought of as an autocorrelation issue. For example, if we take solar wind speed measurements at L1 and split them into training and test data by alternate hours, then both datasets contain observations of the same solar wind streams, CMEs, etc., and the two datasets are not independent. To ensure dependence, the data must be split into chunks that are much longer than any autocorrelations in the time series.

In the revised manuscript, we will add a statement in Section 3.4 ("Data preprocessing") explaining that we use chronological splitting instead of random splitting to avoid data leakage, and cite appropriate references such as Lones (2024) and Apicella (2025).

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.

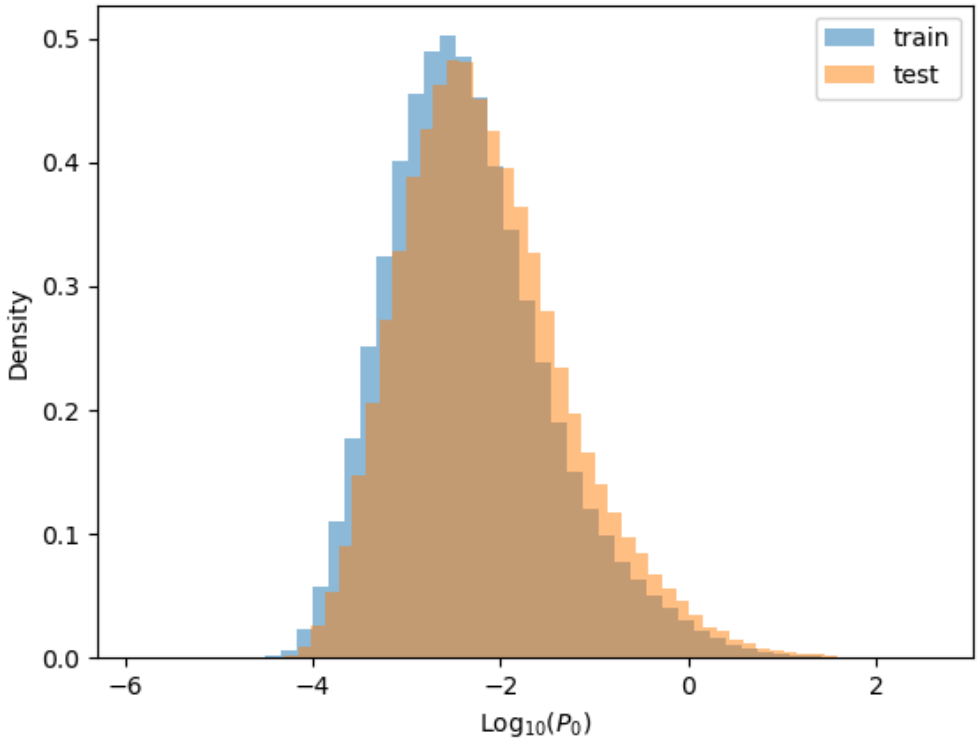

Figure 1. Distribution of logarithm of P0 values for training (blue histogram) and test (orange histogram) datasets

- **The figure below shows the distribution of the training data set (blue) and test data set (orange) after taking the logarithm. It is clear that the test data set, which contains more extreme storms, is shifted to slightly higher values as anticipated by the reviewer: The median, mean, and standard deviation of the training data set are respectively 0.0037, 0.082, and 1.22 in $(nT/s)^2$. For the test data set these are 0.0056, 0.14, and 1.78 in $(nT/s)^2$.**

  While there certainly are statistical differences, as the reviewer points out, we still believe it is most appropriate to split the data as we have (i.e., chronologically) given the high risk of data leakage with random splitting, as described in our reply to the reviewer's comment above.

- **Regarding the reason why the model seems to predict values with reasonable magnitudes, we attribute this to our use of robust scaling and training the model using the base-ten logarithm of P_0 (spectral power).**

- **We propose to raise the point the reviewer makes about statistical differences between the training and test data sets in Section 3.4 ("Data preprocessing") of the revised manuscript. In particular, we propose to add the following paragraph to Section 3.4:**

  An important difference between the training and test data set is that the training data set covers the beginning of the ascending phase of Solar Cycle 24 through the beginning of the ascending phase for Solar Cycle 25, while the test data set only represents the ascending phase of Solar Cycle 25. The occurrence rate of CMEs is higher in the test data set than in most of the preceding solar cycle, and the storms in the test data set are also generally more extreme. This difference is manifest in statistics of P_0*, where for example the mean value of* P_0 *for the training and test data set is respectively 0.082* $(nT/s)^2$ *and 0.14* $(nT/s)^2$*. One seemingly intuitive way to make the training and test data sets more similar statistically would be to split the data randomly. Random splitting of time series data, however, introduces data leakage issues (Lones, 2024; Apicella et al, 2025). We therefore chose to split the data chronologically.*

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).

- **Thank you, we propose to revise this sentence so that it reads as follows:**
  "As L1 is ~1.5 x10^6 km from Earth, such measurements provide lead times of ~30–60 min depending on solar wind speed, which ranges from approximately 300 km/s to 800 km/s."

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."

- **We agree that this ought to be clarified, and we like the reviewer's recommended revision. We propose to incorporate this in the revised manuscript.**

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?

- **The value is chosen randomly. We propose to add the following text following the statement regarding the standard deviation: "The exact value of the standard deviation for a particular arrival time distribution is randomly sampled from a uniform distribution *U*(5 h, 6 h), where this distribution was selected based on typical arrival time distribution widths observed in the HUXt model (Barnard et al, 2022)."**

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?

- **Yes, only the time history of this variable is an input. This is illustrated in Figure 4, though the reviewer points out that this was unclear in the original manuscript. We propose to add a footnote in Table 1 to state that only the time history of log(P_0) is used in model training, as illustrated in Figure 4.**

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?

- **We agree, we propose to revise this statement to simply read that *m* is the number of observations.**

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

- **Thank you for catching this. We propose to revise this equation to indicate that the sum is over *n*, and to modify the text that follows so that it reads "where *f*(*t*) is the forecast value and *f*(*t*-*h*) is the last observed value. Given an averaging window of width *w* and *n* = 0, 1, 2, ..., *w*-1, Equation 4 represents a rolling mean of the previous *w* observations including the most recent observation."**

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.

- **The dropout probabilities were zero, and were indeed tuned using the Optuna search. Thank you for noting that we did not mention this. In accordance with this comment and a comment from Reviewer 2, we propose to modify the description of hyperparameters selected via Optuna in the revised manuscript so that it reads:**

  "The tuned parameters had a batch size of 70, a look-back window Δt_b = 34 h (or N_b = 204 as previously mentioned), an optimizer with a learning rate of 5.4e-05, early stopping at a patience value of 8, and a dropout probability of zero. (The patience value is a hyperparameter used with the early stopping technique; it is the number of epochs the training process must wait for an improvement in the model's performance on the validation set before stopping the training. The dropout probability is the probability of ignoring a random neuron and its connections; it is part of a commonly used strategy for avoiding overfitting that is known as the dropout method (Srivastava et al, 2014).)"

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.

- **Thank you for pointing out that we failed to describe the performance of the persistence and climatology models in isolation. We propose to revise the statement made in the Results section about a "recurrence model" as follows, so that consistent language is used and to describe these additional results:**

  In addition to the benchmark model, we also examined the performance of a simple persistence model (not shown) in which the current value was predicted to be the same as the value measured at an earlier time (e.g., w *= 1 and* h *= 144 in Equation 4 for a persistence model based on measurements made 24 h earlier with data at 10-min resolution). These models yielded coefficients of determination* R^2 *= (85.7, 54.0, 31.2, 16.1, 2.0, −44.0) for forecasting horizons of* h *=(10 min, 1 h, 3 h, 6 h, 12 h, 24 h). We also performed this calculation for the climatology model described in Section 3.3 and obtained* R^2 *= −2.9. In summary, the persistence and climatology generally performed much worse than Models A–C. The exception was the persistence model with a forecasting horizon of 10 minutes, which performed approximately as well as Models A–C.*

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.

- **We propose to modify Table 3 so that it shows the values of TP, FP, TN, and FN as a percentage of the total number of measurements in the test data set (N=153,992); to include the number of positives (2310) and negatives (151682) in the table caption; and to expand the description to include the other parameters.**

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.

- **We agree that this could be made clearer. We propose to revise this statement so that it reads as follows in the revised manuscript:**

  Previous studies have shown that deep learning models generally perform better than conventional machine learning models (e.g., neural networks and nonlinear models; see Keesee et al., 2020; Zewdie et al., 2021; long et al., 2022; Hu et al., 2023, and references therein).

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 Bz, 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.

- **Following the reviewer's recommendation, we propose to make the discussion of these points clearer in the revised manuscript by replacing the previously indicated paragraphs with the following:**

  While Figures 9c–e indicate that CME arrival time is an important input feature for time horizons h ≥ 3 h*, Figure 8 clearly demonstrates that Model C, which is constructed without CME arrival time information, performs just as well as Model A. We therefore conclude that high-resolution solar wind and IMF forecasts are necessary to achieve the model performances we have reported for Models A and C. CME arrival time information is apparently unimportant when high-resolution forecasts of IMF and solar wind conditions are available.*

  In contrast, it is somewhat unclear what role CME arrival time information plays when only low-resolution (but nevertheless accurate) solar wind and IMF forecasts are provided to the model: According to Figures 9c–e, CME arrival time is the single most important piece of information for Model B's predictions. To test this we trained another version of Model B (low-resolution space weather forecast inputs) in which information about IMF Bz and IMF By was excluded but CME arrival time was retained. In this model (not shown), CME arrival time had no importance for any of the five forecast horizons; this model also performed generally much worse than Model B for forecasting horizons h = 6 h and h = 12 h. We therefore conclude that when only low-resolution solar wind and IMF forecasts are available, information about CME arrival time could contribute to model performance. This result incidentally highlights that a study of model performance as a function of temporal resolution and accuracy of solar wind forecast parameters would likely provide clarity around this point.

  To summarize, our results indicate that when only low-resolution forecasts of IMF and solar wind conditions are available, CME arrival time information may improve model performance. In the absence of any information about future IMF conditions, CME arrival time is apparently unimportant. This conclusion corresponds to the qualitative but well known idea that the geoeffectiveness of a CME is closely tied to the accompanying IMF conditions (e.g., Kang et al., 2006).

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"?

- **We agree with the reviewer that the sentences on these lines were unclear. Since this paragraph mostly repeated (or attempted to repeat) previously stated conclusions, we propose to remove it from the revised manuscript.**

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

- **Thank you for noting this omission. These metrics are calculated over the entire test data set. We propose to revise the first mention of Figure 8 in the revised manuscript so that it reads:**

“Figure 8 summarizes performance metrics of each model for each of the five forecasting horizons, calculated over the entire test data set.”

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

- **Thank you for pointing out these references. We propose to add the following parenthetical statement to the conclusion of the paper: “***(It is not, however, the first machine learning-based demonstration of the importance of IMF B_z ; see, e.g., Lotz and Cilliers, 2015; Coughlan et al., 2025.)*”

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.

- **We will make all of these technical corrections according to the referee’s suggestion in the revised manuscript.**