== Title ==

== Team ==

Date: May 5th, 2025

Dr. Pickle Von Mustard
Chief Sandwich Engineer
International Institute of Lunchtime Studies
42 Bread Roll Avenue, Snacktown, ZZ 99999
pickle@mustardmail.fake

Captain Waffletron
Syrup Deployment Specialist
Galactic Pancake Federation
7 Crispy Nebula Crescent, Syrup Sector, M00N-9
waffletron@orbitalsweets.space

Sir Meowington III
Senior Purring Consultant
Institute of Cat-Based Physics
88 Yarnball Street, Catropolis, FLUFF-420
meow3@felineforces.fake

Abstract

The technical reports' scoring scheme assesses the innovation and overall quality of the technical report, including clarity, novelty, technical depth, reproducibility, and insights. The report's score is part of the competition's final composite score. This document contains more details for the finalists who are admitted to Phase 2, and a template that you can use to prepare your report. [a]

[a] All the code can be found in Github: `link`

Guidelines

Deadline

The deadline for submitting your technical report is May 26, 2025, at 11:59:59 PM EDT. Please send your report as a single PDF to `ai_challenge@mit.edu`.

Example of reports

Three examples of reports from the last year Challenge can be found in Sections 3.1, 3.2, and 3.3 here.

Final composite score

Here are some clarifications on the scoring for Phase I and Phase II of the competition. We refer to the Phase I score as the model score (M), which is calculated as follows:

$$M = 0.85 \cdot M_{public} + 0.15 \cdot M_{private} \tag{1}$$

with M_{public} the public weighted score (obtained from the public dataset), and $M_{private}$ the private weighted score (obtained from the private dataset).

M_{public} is based on the orbital decay root mean squared error (OD-RMSE), a weighted skill score that compares the model's RMSE to a baseline

$$\text{OD-RMSE} = \frac{\sum_t \text{Weight}(t) \cdot \left(1 - \frac{\text{RMSE}_{\text{test},t}}{\text{RMSE}_{\text{MSIS},t}}\right)}{\sum_t \text{Weight}(t)} \tag{2}$$

with:

$$\text{Weight}(t) = e^{-\gamma t}, \quad \text{with } \gamma = -\frac{\ln(\epsilon)}{T} \tag{3}$$

where t is the time since the beginning of the evaluation period, γ is the decay constant, ϵ is a small threshold (e.g., 10^{-5}), and T is the total evaluation duration. This metric weights earlier time steps more heavily using exponential decay. Values range from $-\infty$ to 1, where 1 indicates perfect performance and values ≤ 0 indicate no improvement or worse than the baseline.

$M_{private}$ is based on the solar storms-aware OD-RMSE (SSAOD-RMSE), which is calculated as follows:

$$\text{SSAOD-RMSE} = \frac{\sum_t \text{Weight}(t)\,\text{Skill}(t)}{\sum_t \text{Weight}(t)} \tag{4}$$

with the skill at each lead time:

$$\text{Skill}(t) = 1 - \frac{\text{RMSE}_{\text{pred},t}}{\text{RMSE}_{\text{MSIS},t}} \tag{5}$$

and the combined exponential weight:

$$\begin{aligned} \text{Weight}(t) &= e^{-\gamma t} \cdot \left[e^{0.5\,\text{Kp}(t)}\right]^2 \\ \gamma &= -\frac{\ln(\epsilon)}{T} \end{aligned} \tag{6}$$

where t is the time elapsed since the start of the evaluation, $\text{Kp}(t)$ is the planetary kp index for the 3-h segment containing t, $\epsilon = 10^{-5}$ fixes the minimum time weight, and T is the total evaluation duration. Like OD-RMSE, this score favors early lead–times, but it also emphasizes intervals of strong geomagnetic activity (high Kp). The score ranges from $-\infty$ to 1, with 1 indicating perfect predictions, 0 parity with the MSIS baseline, and negative values worse than the baseline.

In addition to the model score from Phase I, we provide the normalized model score, where 1 represents the highest M achieved. The final composite score (CS) is a combination of the normalized model score (M_{norm}) and the report score (Q) for the technical write-up. The composite score is calculated as follows:

$$\text{CS} = 0.8 \cdot M_{\text{norm}} + 0.2 \cdot Q \tag{7}$$

Q assesses the innovation of the approach and the quality of the report, considering its clarity, novelty, technical depth, reproducibility, and insights. The report should not exceed 4 pages in length, including figures but excluding references. The report will be reviewed by a panel of judges and the final score will be the average of the scores given by the judges (Q_j), divided by 25 (the maximum achievable score):

$$Q = \bar{Q}_j / 25 \tag{8}$$

The report score is calculated by summing the scores across the five metrics, with a maximum achievable score of 25, as each metric contributes 5 points. The report will be evaluated by multiple judges, each providing a different set of scores. The total report score for each judge will be summed up, and then an average of the report scores will be calculated.

Consider the following example with 2 judges and a M_{norm} score of 0.95:

- Judge 1: (Clarity) 5 + (Novelty) 4 + (Technical Depth) 5 + (Reproducibility) 5 + (Insights) 4 = 23
- Judge 2: (Clarity) 5 + (Novelty) 4 + (Technical Depth) 4 + (Reproducibility) 5 + (Insights) 4 = 22
- $Q = \bar{Q}_j/25$ =((23+22)/2) / 25 = 0.9
- CS = 0.8 x 0.95 + 0.2 x 0.9 = 0.94

Judging criteria

Here is a detailed breakdown of the judging scheme that will be used to evaluate your technical report:

Clarity (0-5):

- 0: The report is extremely unclear, with incomprehensible language and disorganized content. It lacks coherence, making it nearly impossible to extract meaningful information. Sentences may be convoluted, and transitions between sections are absent.
- 1: The report is very unclear, with frequent use of jargon and poor organization. Readers struggle to follow the narrative, leading to confusion. Key points may be buried within dense paragraphs.
- 2: The report is somewhat unclear, with occasional use of jargon and a lack of concise explanations. The organization could be improved for better readability. Some sections may require restructuring to enhance clarity.
- 3: The report is generally clear, with minimal use of jargon and adequate organization. However, some sections may still benefit from further simplification. Sentences are straightforward, but occasional complexity persists.
- 4: The report is clear and concise, making it easy to follow. Jargon is absent, and the content flows logically, showcasing excellent organization. Each paragraph serves a purpose, contributing to the overall coherence.
- 5: The report is exceptionally clear, well-structured, and easy to understand. It is devoid of jargon and outstandingly organized, ensuring effortless understanding. The writing style is elegant, and transitions between ideas are seamless.

Novelty (0-5):

- 0: The approach lacks any originality, merely rehashing existing knowledge without introducing anything new. It fails to innovate and falls short of making a meaningful impact.
- 1: The approach is largely unoriginal, offering little in terms of innovation. It follows established methods without introducing novel concepts, limiting its potential for impact.
- 2: While displaying some creativity, the approach lacks significant originality. It introduces a few unique elements but does not revolutionize the field. Incremental improvements are noticeable.
- 3: The approach shows some novelty, introducing fresh ideas or techniques. It demonstrates creativity, although without groundbreaking impact. It may adapt or combine existing methods in interesting ways.
- 4: The approach is quite novel, incorporating several new ideas or techniques. It stands out, pushing boundaries and capturing attention. The synthesis of different concepts leads to promising results.

- 5: The approach is highly innovative, introducing groundbreaking concepts. It disrupts the status quo, making a significant impact on the field.

Technical Depth (0-5):

- 0: The report lacks technical detail, offering no explanation of methodologies or approaches. It fails to provide insights, leaving readers uninformed and frustrated.
- 1: Minimal technical detail is provided, with basic concepts briefly mentioned but not elaborated upon. Key aspects remain unexplored, limiting the depth of understanding.
- 2: The report includes some technical detail, touching on methodologies without delving deep. While readers gain a basic understanding, they are left wanting more detailed explanations.
- 3: A moderate level of technical depth is present, explaining most methodologies and approaches. Readers acquire a solid understanding, although some nuances may be overlooked.
- 4: The report is highly detailed, meticulously explaining methodologies, approaches, and technical aspects. It provides thorough coverage, satisfying readers' curiosity and enhancing clarity.
- 5: An extremely detailed report, providing comprehensive explanations. Every methodology, approach, and technical facet is thoroughly covered, making it a valuable resource for advanced readers.

Reproducibility (0-5):

- 0: The report lacks any mention of reproducibility or provides no information for replication. It lacks transparency, hindering others from reproducing the reported results.
- 1: While the report mentions reproducibility, it does not offer sufficient information for replication. Key details are missing, making it challenging for others to recreate the experiments.
- 2: The report provides some information on reproducibility, but it lacks clarity in describing how to replicate the results. Certain steps or parameters remain ambiguous.
- 3: The report is somewhat reproducible, providing sufficient information for replication, but it could be more explicit in certain areas. Some implementation details or data preprocessing steps may benefit from further elaboration.
- 4: The report is highly reproducible, offering clear descriptions and explanations for most aspects of the model's reproducibility.
- 5: The report is extremely reproducible, providing exceptionally clear descriptions and explanations for all aspects of the model's reproducibility. A diligent reader can precisely replicate the reported results.

Insights (0-5):

- 0: The report lacks any meaningful insights and does not draw any valuable conclusions. It merely presents facts without interpretation.
- 1: The report provides minimal insights, failing to draw valuable conclusions. It summarizes findings without delving into their significance.
- 2: While the report provides some insights, it could be more insightful and impactful. Conclusions are surface-level, lacking depth.
- 3: The report offers moderate insights, presenting valuable conclusions, implications, and lessons learned. It connects findings to broader contexts.
- 4: A highly insightful report, providing valuable conclusions, implications, and lessons learned. It synthesizes results, identifies patterns, and suggests practical applications.
- 5: The report reaches exceptional insights, offering profound conclusions, implications, and lessons learned.

Report Template

This template provides a suggested structure for your report, but you are welcome to use alternative individualized sections or narratives, and any writing software you prefer. While not required, we strongly recommend using LaTeX, as the top three finalists will be invited to contribute to a joint conference paper similar to last year's Challenge [1].

Introduction & Approach Definition

This section should clearly articulate your chosen strategy, outlining your conceptual workflow, key assumptions, and any innovative elements. Explicitly state whether your approach is data-driven, physics-informed, or a hybrid, justifying your choice and referencing relevant background literature to contextualize your approach, e.g. [2].

Data & Pre-processing

This section should provide a thorough description of the data sources used, including their coverage, quality control procedures, and any transformations applied. For example, you can include a visual representation (e.g., a plot) of the raw and cleaned data distributions to illustrate the impact of your pre-processing steps on the data.

Methods

This section should detail your model design, training protocol, and any baseline models used for comparison, presenting a clear architectural diagram and hyperparameter-response curves where relevant. For example, you can explain how you trained your model, the algorithm, the metrics, and dataset splitting.

Results

This section should present key performance metrics to quantify your model's performance, e.g. including lead-time specific results and uncertainty quantification. You can use multi-panel plots to compare prediction error as a function of lead time and to visualize event-based detection rates.

Discussion & Model Evolution

This section should analyze the limitations of your current model and propose specific ways it could be adapted or evolved in future work, e.g. discussing the potential for incorporating continual learning techniques, new data streams, or physics-based feedback loops.

Conclusion & Broader Impact

This section should summarize the key takeaways from your work, highlighting the operational implications of your model and any relevant considerations. Emphasize how your evolving model can improve real-time forecasting capabilities and contribute to mitigating the risks associated with space weather events.

Code availability

This supplementary section should provide details on accessing the software and ensuring the reproducibility of the results.

References

[1] P. M. Siew *et al.*, "Satellite pattern-of-life identification challenge: Competition design and results," *25th Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS)*, Sep. 2024.

[2] J. Briden, P. M. Siew, V. Rodriguez-Fernandez, and R. Linares, "Transformer-based atmospheric density forecasting," *Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference*, 2023. [Online]. Available: https://arxiv.org/abs/2310.16912.