Accuracy analysis of the CME predictions for 2024

Marina Serra Cabra

Contents

Abstract

This report touches upon the research made within my internship at GeoSphere. It especially focuses on the analysis of the Coronal Mass Ejections (CME) predictions made by the community (CCME) for 2024 and compares them to recorded local and global data in order to prove their accuracy in both scales. Despite there are several arrangements that need to be made in the future, this analysis has led to the conclusion that the actual values in Kp are generally lower than predicted and that the CME predictions' accuracy in local and global terms is less than 40 %.

1 Introduction

Space weather, which is the term used to describe the conditions generated by solar wind's, Earth magnetic field's, and ionosphere's coupling (Pulkkinen (2007)), is a topic of great relevance nowadays. The main reason for this is the fact that human technology has expanded towards the region of space in which the consequences of space weather, especially the coronal mass ejections (CME), which are events that produce solar wind, affect the systems, thus putting them at risk of damage (Haberle (2023)).

Many different approaches can be taken to sort this problem, one of them being CME predictions, which could enable governments or other institutions to implement measures before any remarkable event takes place. This is why this report focuses on analyzing the accuracy of the CME predictions recorded by the Community Coordinated Modeling Center (CCMC) CME Scoreboard from NASA (Wiegand (2024)) during 2024, both on a local and a global scale.

In order to do so, local and global data need to be considered. This project uses data from the Conrad Observatory in Austria as local (its K index), as well as some other indices that contain global data. These are the K-planetary (Kp) and the Disturbance Time (Dst) indices. In the Kp index's case, the K-indices and Ks-indices from 13 geomagnetic observatories are combined, with the vast majority of them being found in mid-latitudes and over Europe (Haberle (2023)). On the other hand, the Dst index is derived from four observatories located in low latitudes, coinciding with the placement of the ring current, which induces depression of the horizontal magnetic component (Haberle (2023)).

Within the following sections, the methodology, results, discussion, and conclusion of this project are presented. The methodology section contains an accurate description of the tools and the data used for the analysis, as well as some important guidelines that were followed. The results and discussion section provides the final outcome in the form of pie charts and skill scores, reflecting on it at the same time. Finally, the conclusion outlines the main contribution of this project to the scientific community as well as suggestions for future improvement to tackle the issues encountered.

2 Methodology

The data analysis was conducted using Python in a jupyter notebook [1] and inside a pandas environment in order to facilitate the process of a data frame creation.

The data used for the analysis were extracted from Nose et al. (2015) for the Dst, Matzka et al. (2021) for the Kp and Team (2024) for the K for 2024. In the following subsections, an accurate explanation of the methodology followed to process each of the data needed for the analysis (K, Kp and Dst) is presented.

2.1 Kp and K analysis

In the Kp and K case, the data was already given in numbers in three-hour time ranges, which had to be selected to match the CME prediction arrival date. This was achieved by using a "while loop" in the Python programming options that would go through all the K and Kp data and stop when the date and time would be the closest to the CME prediction arrival date and time, always looking from the past. Then, two columns containing the actual K and Kp values were created to match the value found in the Kp and K data frames to the CME predictions.

After that first preparatory step, the Kp column in the CME prediction was compared to the Kp and K actual values assigned from the formerly mentioned data frames. In order to achieve this, two more columns and further programming were required. These new columns were called 'InRangeGlobal' (for the Kp comparison) and 'InRangeLocal' (for the K comparison) and were filled with the values 'lower' when actual values were lower than predicted,

[1] The jupyter notebooks used are attached to this report.

'higher' when actual values were higher than predicted, and 'yes' when the actual value was in the predicted range.

To conclude with the K and Kp analysis, CME arrival time was also taken into account. For this, two more columns were created and filled manually using the data plot from Figure 1, made by plotting all the data from the predictions along with the Dst (which was not used in this case, but later on, see 2.2). These were named 'timingK' and 'timingKp' and were filled following the contingency table presented in Figure 2, initially presented as Figure 2 in Verbeke et al. (2019). However, it needs to be noted that, due to the focus on only the predictions and the storms surrounding them, the 'miss' value was not defined the same way as in the contingency table but was defined as a closer-in-time false alarm, as stated below.

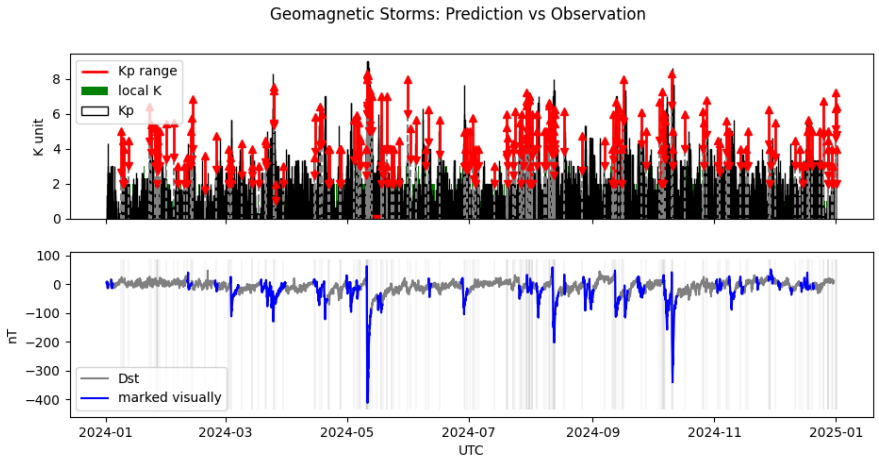

Figure 1: CME predictions (red arrows) plotted along with the K and Kp rates (green and white bars) in the upper graphic, and Dst flagged index in the graphic below (see 2.2.2 for more details on Dst flagging)

	Observed Arrival	No Observed Arrival
Predicted Arrival	Hit (H)	False Alarm (FA)
No Predicted Arrival	Miss (M)	Correct Rejection (CR)

Figure 2: Contingency table showing the categorization used when classifying different predictions based on their local (K) and global (Kp and Dst) accuracy.

The values inside the columns included information about whether the prediction was made without an actual event happening (FA, standing for 'False Alarm'), the actual event happened at a different time not further away than 24-48 hours before or after the prediction (M, standing for 'Miss'), or the actual event happened right at the same time as the predicted time of arrival (H, standing for 'Hit'). Correct rejection was out of this research's scope.

In the discussion section, four pie charts plotted using the data from the 'InRangeGlobal', 'InRangeLocal', 'TimingKp', and 'TimingK' columns are presented and taken into account as a remarkable feature that predictions should cover accurately (see sections 3.1 and 3.2).

2.2 Dst analysis

The analysis of the Dst index was developed by performing a previous procedure, which involved flagging all storms for 2024 manually using the xmagpy tool.

After that process, which is described in 2.2.1, the timing in the flagging and the CME prediction times were compared in order to take that into account once again in the discussion. The Dst comparison methodology is described in Section 2.2.2.

2.2.1 Solar storm flagging

The Dst graphic of the 2024 curve was flagged using xmagpy (see the graphic at the bottom of Figure 1). Some relevant considerations need to be taken into account when it comes to marking solar storms manually. In this case, a solar storm's start was characterized by the Storm Sudden Commencement (SSC), which is when the function starts rapidly growing (see Figure 3),

and its end was matching the end of the recovery phase, which is when the function gets to zero again (once again, see Figure 3).

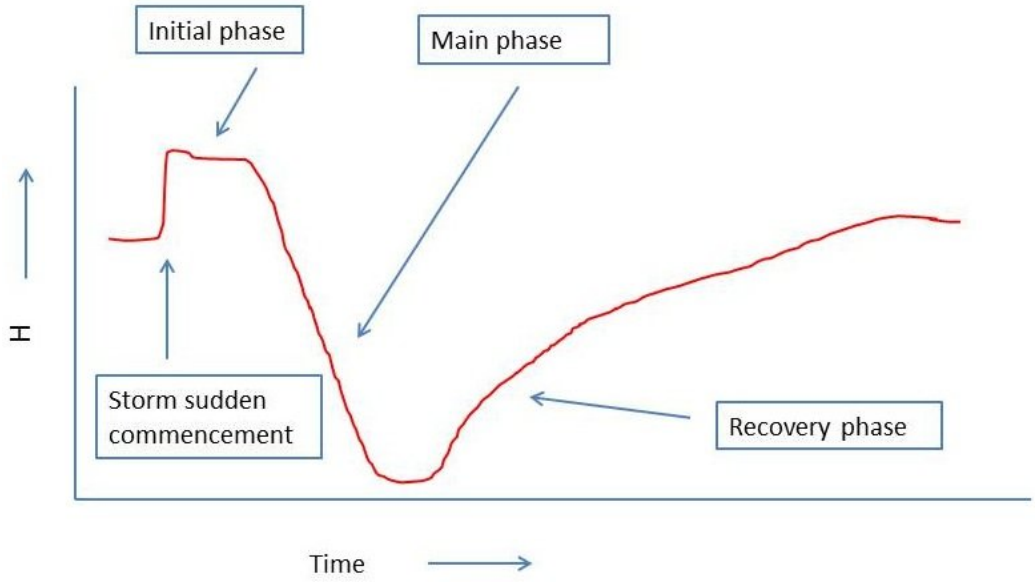

Figure 3: Parts of a solar storm

These flags were saved as a file and processed via Python into a pandas data frame that contains the start and end date for each storm.

2.2.2 Data treatment

The CME predictions were compared to the Dst index flags, obtained as section 2.2.1 details, using the same contingency table shown in Figure 2 and the same specification for the 'miss' values as before but with different time ranges, detailed below.

This implies that the 'OnsetDst' column, created specifically for the comparison mentioned previously, included information about whether the prediction was made without an actual event happening (FA, standing for 'False Alarm'), the actual event happened at a different time not further away than 12 hours before or after the prediction (M, standing for 'Miss'), or the actual event happened 1 hour before or after the prediction, or even at the same time (H, standing for 'Hit'). Correct rejection was, as with the K and Kp ranges and timing, beyond the scope of this data analysis.

In the discussion section, a pie chart plotted using the data from the 'OnsetDst' column is presented and taken into account as a remarkable feature that predictions should cover accurately (see section 3.2).

3 Results and discussion

As stated in sections 2.1 and 2.2.2, the information from the remarkable columns in the pandas data frame that was created to proceed with the analysis of CME predictions' accuracy was plotted into pie charts to be analyzed more in depth and is presented within the following sections.

In addition, several relevant skill scores for this investigation were calculated using the formulas presented in Table 9 in Verbeke et al. (2019) (see Figure 4). These are the hit rate (see section 3.3) and the accuracy (see section 3.4). Note that no other rates from the ones presented in Figure 4 could be calculated due to the slight distinction made in the 'miss' parameter (see section 2.1) and to not having a correct rejection rate. For instance, the success ratio and the false alarm ratio could not be calculated because the 'miss' value should also have been taken into account, which is what is already happening within the accuracy rate, also changing a bit its definition in this investigation's scope. On the other side, the false alarm rate and hence the Hanssen & Kuipers score would not make sense because, without the correct rejection, the score would become lower than 0, leading to a greater difficulty in interpretation.

Skill score	Equation	Perfect score	Comments
Hit rate (POD)	$\frac{H}{H+M}$	1	Fraction of observed arrivals that were predicted.
Success Ratio (SR)	$\frac{H}{H+FA}$	1	Fraction of correct predicted arrivals. False Alarm Ratio = 1−SR
Bias Score	$\frac{H+FA}{H+M}$	1	Ratio of predicted arrivals to observed arrivals, <1= underforecast; >1= overforecast
Critical Success Index (CSI)	$\frac{H}{H+M+FA}$	1	Fraction of correct observed arrivals.
Accuracy	$\frac{H+CR}{Total}$	1	Fraction of correct forecasts.
False Alarm Rate (POFD)	$\frac{FA}{CR+FA}$	0	Fraction of incorrect observed nonarrivals
Hanssen & Kuipers discriminant	$HK = POD - POFD$	1	Forecast ability to discriminate between observed event occurrence from nonoccurrence

Figure 4: Derived skill scores from contingency table. Extracted from Verbeke et al. (2019).

3.1 Local prediction

As stated in section 1, the relevance of this research lies in the fact that not only CME predictions are analyzed in order to prove their accuracy, but this accuracy is compared in local and global terms.

In this section, a comparison between the local K values and the predicted values for CME is presented in the form of two pie charts: one for the value and one for the timing. The advantage of presenting the results like this is that it is easy to quantify them using a percentage, which is helpful when talking about accuracy. Further on, as stated in the beginning of section 3,

some rates related to these comparisons are calculated (see sections 3.3 and 3.4).

The first chart (see Figure 5) shows whether the actual K was in the predicted range, lower, or higher. As can be seen, most of the actual K values are lower than predicted. This suggests that the CME predictions are not that accurate in local terms; therefore, some arrangements should be made from each local institution over the general prediction in order to get more exact ones.

On its right side (see Figure 6), the timing K values are presented. As stated in the methodology, the chart shows the percentage of the values 'H' (hit), 'M' (miss), and 'FA' (false alarm). In this case, the number of hits and misses are notorious, which means that the predictions on a local scale were quite accurate in time (remember that a miss implies a 12-hour difference in the future or in the past at its most). However, since only half of these predictions are hits, there is still room for progression in regard to the exact timing, which could also be made, as stated before, by a filtering for each region regarding its coordinates.

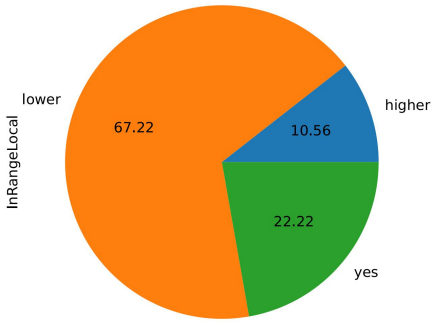

Figure 5: K values analysis

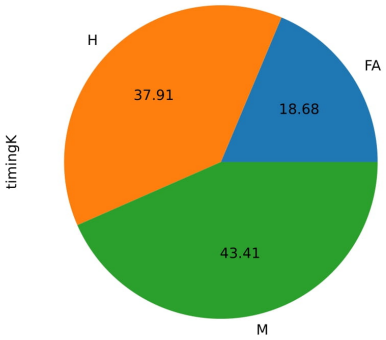

Figure 6: K timing analysis

3.2 Global prediction

In this section, a comparison between the global Kp values and the predicted values for CME is presented in the form of two pie charts, along with a pie

chart for the Dst timing analysis. Further on, as stated in the beginning of section 3, some rates related to these comparisons are calculated (see sections 3.3 and 3.4).

The first chart (see Figure 7) shows whether the actual Kp was in the predicted range, lower, or higher. As can be seen, most of the actual Kp values are lower than predicted, even more than in the K case. In addition, the 'yes' percentage is also smaller (16.48 % compared to the K chart's 22.22 %). This suggests that the CME predictions, contrary to the first assumptions in this research, are even less accurate on a global scale.

That might be due to the geographical coordinates of the observatories collecting data for the Kp index, which are in relatively mid-latitudes, leading to lower Kp values compared to a global average. However, regarding the fact that they are located in the mid-latitudes, which should be where the average predicted rate is detected, a systematic error in the predicted rates for the Kp could also be a cause of these results. The majority of CME actual Kp values being lower than predicted (71.53 %) is another strong argument for the systematic error that predictions could be taking for granted.

On the right side of the previously discussed chart (see Figure 8), the timing Kp values are presented. As stated in the methodology, the chart shows the percentage of the values 'H' (hit), 'M' (miss), and 'FA' (false alarm). In this case, the number of hits and misses are even more notorious than with the K timing, which means that the predictions on a global scale were quite accurate in time (remember that a miss implies a 12-hour difference in the future or in the past at its most). However, since the percentage (%) of misses has increased and the hit one has decreased compared to the K timing chart (see Figure 6 in section 3.1), it can be stated that the accuracy is overall lower. Once again, this could be a problem related to the geographical positioning of the Kp observatories, which could be the reason for a late or early detection of the CME compared to the prediction. To tackle this problem, more accurate data on whether the misses were later or earlier is needed as a first step.

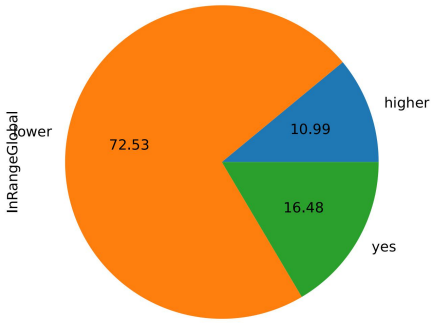

Figure 7: Kp values analysis

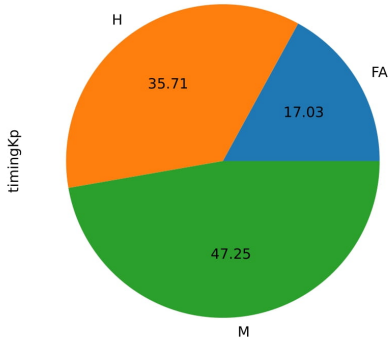

Figure 8: Kp timing analysis

Moving on to the last chart (see Figure 9), the Onset Dst values are presented. As stated in the methodology, the chart shows the percentage (%) of the values 'H' (hit), 'M' (miss), and 'FA' (false alarm). In this case, half of the predictions were a false alarm, whereas 37.63 % were a hit and only 12.37 % a miss. This, compared to the Kp, shows that the predictions of the CME arrival time were generally less in a 12-hour range (50 % were hits and misses in this case compared to the 82.97 % in Kp's timing case). However, they were more accurate when only taking misses and hits into account (more than half of the hits and misses joined together were hits in this case, whereas in the other it was misses that predominated).

The relation here is more complicated, since there are more false alarms but fewer misses, and therefore that cannot be related to the latitude. Hence, further insight into false alarm situations is suggested here, as they account for a significant amount in this graphic (50 %). This implies calculating the time difference between the false alarm and the next or previous storm in order to potentially identify a pattern in the predictions that could have led to these results.

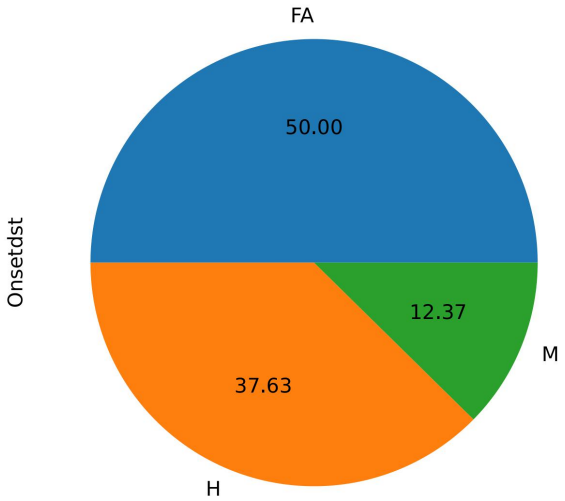

Figure 9: Dst timing analysis

3.3 Hit rate

The hit rate (POD) is a value between 0 and 1 that relates the number of observed arrivals that were correctly predicted to the ones that were observed, as stated in Figure 4 (see section 3). For this reason, it is calculated as follows:

$$POD = \frac{H}{H + M}$$

By doing this calculation, the following values were obtained:

Data	Hit rate	Percentage (%)
K	0.4294478527607362	42.94
Kp	0.4304635761589404	43.05
Dst	0.7526881720430108	75.27

Table 1: Hit rate for K, Kp and Dst predicted timing given in numbers from 0 to 1 and in percentage (%).

By looking at the previous values contextualized with the charts presented in the previous sections (see sections 3.1 and 3.2), it can be stated that, as was already seen in the charts, the Dst has the highest hit rate when comparing it only to the misses, which could suggest that the predictions are made according to or casually matching the Dst observatories' latitudes. Nevertheless, false alarms need to be taken into account, and that makes reaching a conclusion and possible suggestions more complicated (see section 3.4).

On the other hand, the percentage values are in the K and Kp cases below 50 %, for which it can be stated that CME predictions are not accurate, since they get more 'M' than 'H'. A possible explanation is the latitude of both the indices and the systems used in the predictions, so this should be reviewed.

3.4 Accuracy

The accuracy in this case, even though the formula is equal to the critical success index, is defined as the success ratio would be in Verbeke et al. (2019). It takes into account the new definition this report has adopted for the 'miss' term (see section 2.1) and is misregarding correct rejection. Therefore, the formula used is as follows:

$$Accuracy = \frac{H}{H + M + FA}$$

Data	Accuracy	Percentage (%)
K	0.3791208791208791	37.91
Kp	0.35714285714285715	35.71
Dst	0.3763440860215054	37.63

Table 2: Accuracy for K, Kp and Dst predicted timing given in numbers from 0 to 1 and in percentage (%).

Note that this was already calculated in the charts in sections 3.1 and 3.2, where it appears as the hit percentage.

As stated before, the main problem that the data obtained from this project in accuracy terms presents is the low accuracy of the predictions' comparison to the Dst, which had a high hit rate and would be expected to have a high accuracy too. On the other hand, the overall accuracy values

are too low (under 50 %) to continue making predictions without any further arrangement either in the prediction systems or the observatories' data obtention and treatment.

4 Conclusion

This project has focused on analyzing the accuracy of the CME Kp value and arrival time predictions made for 2024 locally and globally using K, Kp, and Dst indices. This has been achieved by creating a pandas data frame with the arrival time and Kp values of all the predictions for 2024 and generating new columns to compare these times and values with the actual ones. The comparison outcome for the values could either be 'yes', 'lower', or 'higher' (see section 2.1 for more details), whereas the result in the time comparison was either 'H', 'M' or 'FA' (see sections 2.1 and 2.2.2 for further insight into the meaning in both Kp and Dst terms). The obtained columns were all plotted into pie charts, and two different skill scores were calculated and discussed. A summary of these discussions will be presented within the following lines, as well as further steps that should be followed in the future.

To conclude with the discussion of the results presented in section 3, it has to be stated that overall, as shown in the accuracy section (see section 3.4), the time predictions have a very low accuracy (under 50 %), and, at the same time, the actual Kp and K values tend to be lower than the Kp predicted range, as presented within the local and global prediction sections (see sections 3.1 and 3.2).

Focusing on the time predictions, from the hit rate in section 3.3, it can be extracted that local K was less commonly hit than global Kp. However, both had a low hit rate, around 40 %, compared to the Dst one, which is around 75 %. That could initially lead to the conclusion that the predictions better adjust to the Dst, which uses data from observatories lying in low latitudes. Nevertheless, the 'FA' proportion shown in the pie charts (see Figures 6, 8, and 9) indicates that the predictions have 50 % of 'FA' when compared to the Dst, while they have around 43-47 % of 'FA' in the local K and global Kp. Therefore, while corrections in terms of latitude could be applied in order to improve the 'H' ranges in local K and global Kp, further investigation and filtering needs to be made in both seeking whether the predictions labelled with an 'M' are from before or after the actual storm and taking further insight into the 'FA' cases in the Dst data.

On the other hand, the Kp value predictions were generally lower, both globally and locally. While that could be due to the fact that the global Kp and local K latitudes (being both mid-latitudes) could both be considered separate from where the greatest values are registered, the next steps that this investigation suggests are that predictions are further analyzed to seek for any systematic error that may lead to these most of the time incorrect predicted ranges.

The previously mentioned outcomes should not be seen as the absolute truth, since, to start with, this research has only focused on arrival times and Kp value ranges. Other parameters such as the velocity or the influence of background solar wind Verbeke et al. (2019) still remain to be investigated in the context of predictions.

In addition, there are some improvements that remain to be made and further considerations that could be the key to clearer conclusions, which are exposed within the following lines.

Firstly, the difference between the contingency table's conception of 'miss' and the one adopted by this analysis should be corrected, since that would enable the calculation of the success ratio, and therefore the false alarm ratio, and would also imply a change in the results and thus their discussion. In order to start with this, all the flagged storms in the Dst index should be taken into account, and the ones without a clear prediction should be classified as a 'miss', rather than classifying predictions without a corresponding solar storm as a 'miss'. To do that, the whole pandas data frame designed for this analysis should need to be redesigned so that all the flagged storms appear in function of the predictions and not the other way round (as it is right now).

Secondly, the manually made processes, such as flagging the storms in the Dst index or classifying the K and Kp timing as 'H', 'M' or 'FA', could have led to errors. Therefore, these should be redone several times, and some parameters, such as the 'H' or 'M' definitions, should be changed in order to minimize that error. For instance, 'H' could also be considered within a 6-hour range instead of a 3-hour one. Since there are no clear time intervals that should be used (Verbeke et al. (2019)), that could also be a valid approach to broaden the research and get better results in accuracy terms.

In third place, as suggested above, in order to gather more data and get to clearer results and thus its interpretation, the 'H', 'M' and 'FA' classification could be more specified. For example, by calculating the ME and MAE of

the time difference, as shown in Verbeke et al. (2019), which would make it possible to state whether the prediction was made before or after the missed event, and could even help find a pattern for the concerning amount of 'FA' in the predictions compared to the Dst.

Getting into formal aspects, the figures could have been plotted so that text does not overlap, and the pandas data frames mentioned within the report could have been exported and added, which was not possible due to a lack of time.

Overall, this investigation has given an answer to the need of evaluating the predictions for solar storms made in 2024. However, as seen above, further tweaking and research on this topic needs to be done.

References

Haberle, V. (2023), Determination of space weather effects on the geomagnetic field : an automatic derivation of geomagnetic baselines containing quiet variations, PhD thesis, IRAP.

Matzka, J., Stolle, C., Yamazaki, Y., Bronkalla, O. & Morschhauser, A. (2021), 'The geomagnetic kp index and derived indices of geomagnetic activity', *Space Weather* **19**.

Nose, M., Iyemori, T., Sugiura, M., Kamei, T., Matsuoka, A., Imajo, S. & Kotani, T. (2015), 'Geomagnetic dst index'.

Pulkkinen, T. (2007), 'Space weather: Terrestrial perspective'.

Team, C. O. (2024), 'Conrad observatorium'.

Verbeke, C., Mays, M. L., Temmer, M., Bingham, S., Steenburgh, R., Dumbović, M., Núñez, M., Jian, L. K., Hess, P., Wiegand, C., Taktakishvili, A. & Andries, J. (2019), 'Benchmarking cme arrival time and impact: Progress on metadata, metrics, and events', *Space Weather* **17**.

Wiegand, C. (2024), 'Cme scoreboard'.