Product: “ap Predictions” **G.171**
This new tool will be a new service of the A.Ne.Mo.S. Cosmic Ray Group of NKUA for the G-ESC centre of the ESA SSA SWE Network.

This page intentionally left blank

Table of Contents

Validation Analysis for product G.171

The validation analysis performed for the most recent version of ap_predictions tool, i.e. 'ap_pred_e10_b4_pack24.h5'.

Introduction

The scope of the tool “ap_predictions” is to forecast the values of the ap geomagnetic index for the next 3, 6, 9, 12, 24, 48 and 72 hours. The development of the tool is based on the state-of-the-art machine learning algorithms and especially the Long-Short Term Memory (LSTM). The training of the model performed on historical data of the 3-hour time intervals of ap index (Jan 1996 – July 2017) and the validation of the tool is performed on unseen data (July 2017 – Nov 2022), i.e. data out of the training interval.

There is no standard set of metrics used by geomagnetic index predictive-model developers to benchmark their models (see e.g. Liemohn et al., 2018). Nonetheless, our goal is to provide the most complete validation analysis performed to date for the ap geomagnetic index. To accomplish this task, we implement a set of variables that are used by various researchers (see e.g. Paouris et al., 2021 and references there in). Furthermore, a very comprehensive analysis of the various metrics can be found in the work of Jolliffe and Stephenson (2012). Our validation analysis is based and is in agreement with the “Guidelines for common validation in the SSA SWE Network”.

Brief Description of the Running Process

The LSTM model uses as input the most recent timeseries of ap index (24 values of 3-hour time intervals) and predict the next one. For example, if the algorithm runs at April 1st at 11:00 (running timestamp: T0) and the last available ap value was the one of April 1st at 09:00 then the algorithm will predict the ap value for April 1st at 12:00 (T1). Then the algorithm is taking into account all the available values (including the predicted one) to forecast the next 3-hour time interval (T2) and this procedure continues up to 72 hours in advance from the running time T0.

Obviously, the algorithm is ready to provide results (in a few seconds) as soon as possible the new ap data are available from the Geomagnetic Observatory Niemegk, GFZ German Research Centre for Geosciences. That implies the ap_predictions tool is capable to provide forecasts almost 3 hours in advance for the first prediction of ap (T1). The leading time is associated with the availability of the data from the GFZ and is independent from the processing time, as the processing time is only a few seconds.

Validation Analysis

For our validation analysis we are utilizing a random set of input timeseries of ap index from the unseen data (July 2017 – November 2022) to predict the next values and then we compare these values to the ground truth as well as with the persistence model (PERS). As the extreme fluctuations of the ap index for two consecutive values are extremely rare the persistence model is expected to perform quiet well and is the most demanding in terms of model comparison. The LSTM forecast model should be able to beat persistence in most measures and metrics (see e.g. Bailey et al., 2022). In our validation analysis we compare LSTM with the PERS models.

For the validation purposes we run each experiment (random selection of input timeseries) 2000 times without any criterion for the period of the solar cycle or if the selected period is described as quiet, active or if it includes a geomagnetic storm. This way is very representative of actual conditions as the opposite will include some bias (e.g. testing on quiet or active periods only). Furthermore, this approach should be considered as the lower limit (worst case scenario) of the performance of our model. In real time mode the algorithm takes into account the influence of Earth directed Coronal Mass Ejections (CMEs) and Earth facing coronal holes (CHs). The associated information for e.g. the Earth directed CME is provided by the space weather (SW) forecaster on duty of the National and Kapodistrian University of Athens in a daily basis. Then the algorithm provides forecasts of ap taking this information into account, so the new values are significantly different than the ones where no active or storm conditions are expected. The important outcome is that the real time tool will have much better metrics than the one which used for the validation analysis as there are no historical data available (from the SW forecaster on duty).

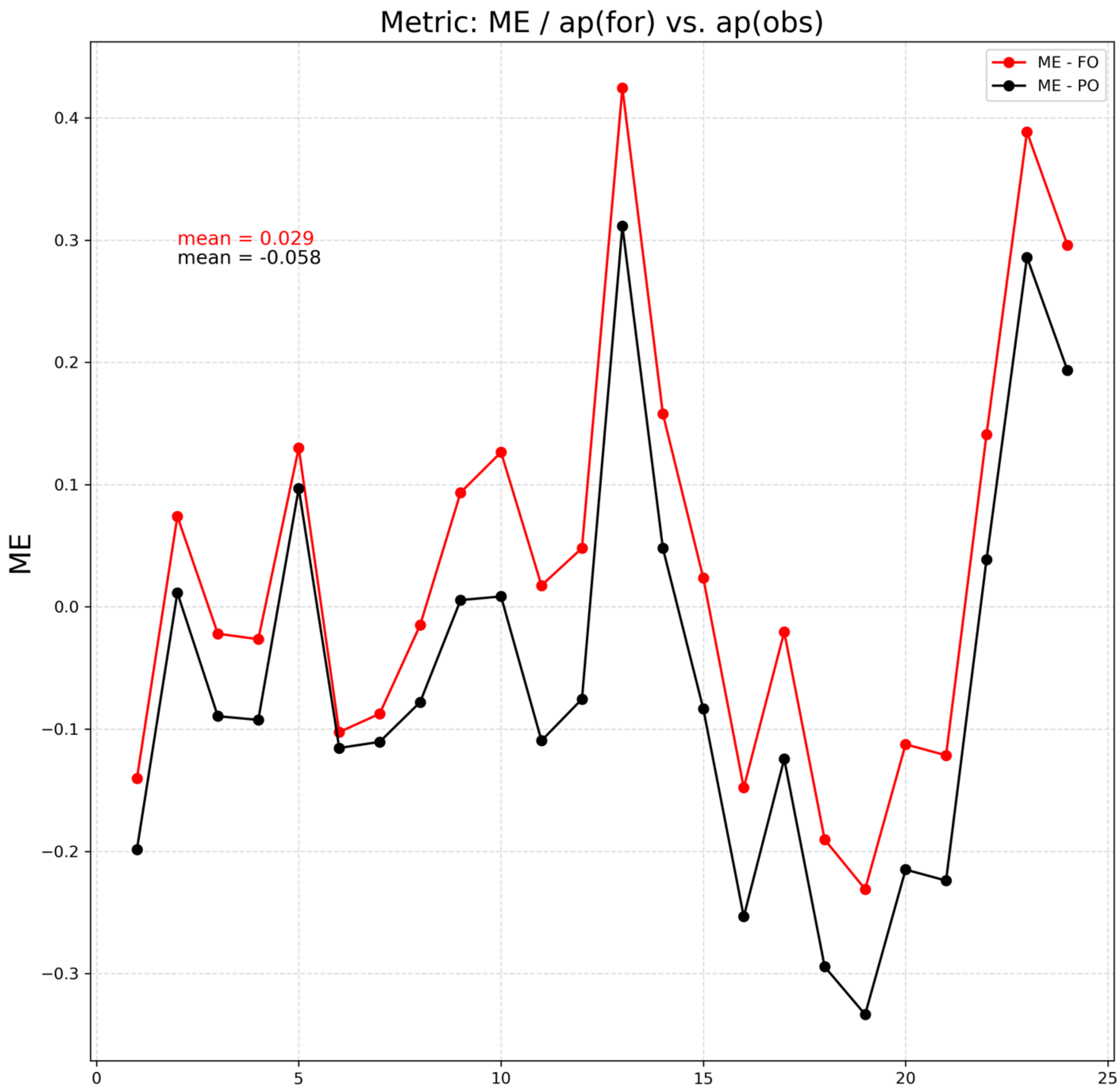
The comparison of the LSTM and PERS models with the ground truth is performed with two different sets of metrics, each one reveals different characteristics. The first category (Category I) is the ***Fit Performance Metrics*** and the second one (Category II) is the ***Threshold Performance Metrics***.

A. Fit Performance Metrics

At this category, the performance metrics are: the mean error (ME) [0], the mean absolute error (MAE) [0], the root mean squared error (RMSE) [0], and the bias (BIAS) [1]. The value corresponds to the best performance is inside the brackets.

ME

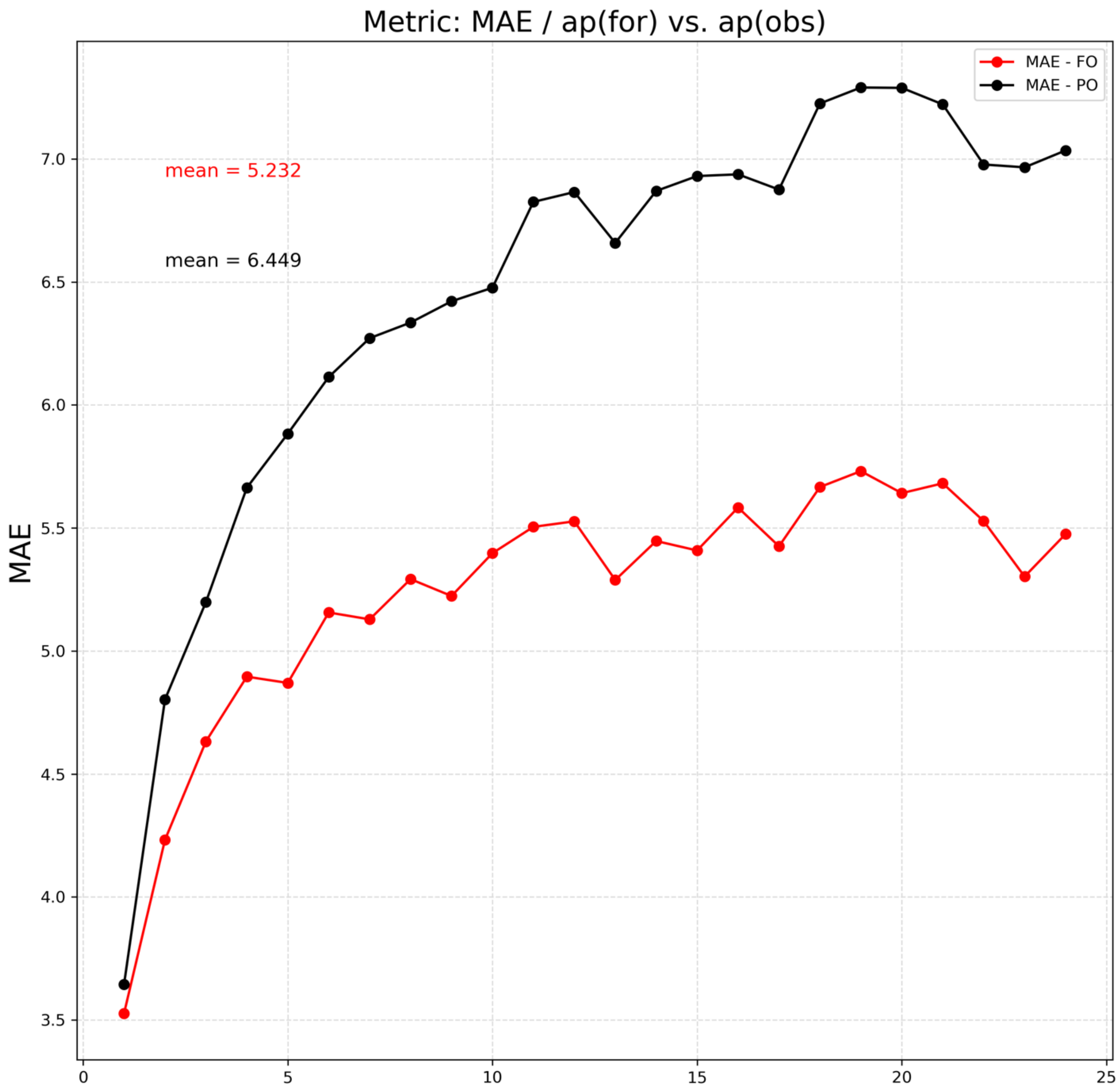
The best performance of the model is for a ME = 0.0. In the next figure we see that LSTM model has a better performance than the PERS model. In 15/24 (62.5%) forecast time windows the LSTM has a ME closer to 0 than the PERS model (37.5%). The mean value for LSTM and PERS models is 0.029 and -0.058 respectively implying that LSTM model is closer to the real values of ap.



ME as a function of time intervals for LSTM model (red) and the PERS model (black).

MAE

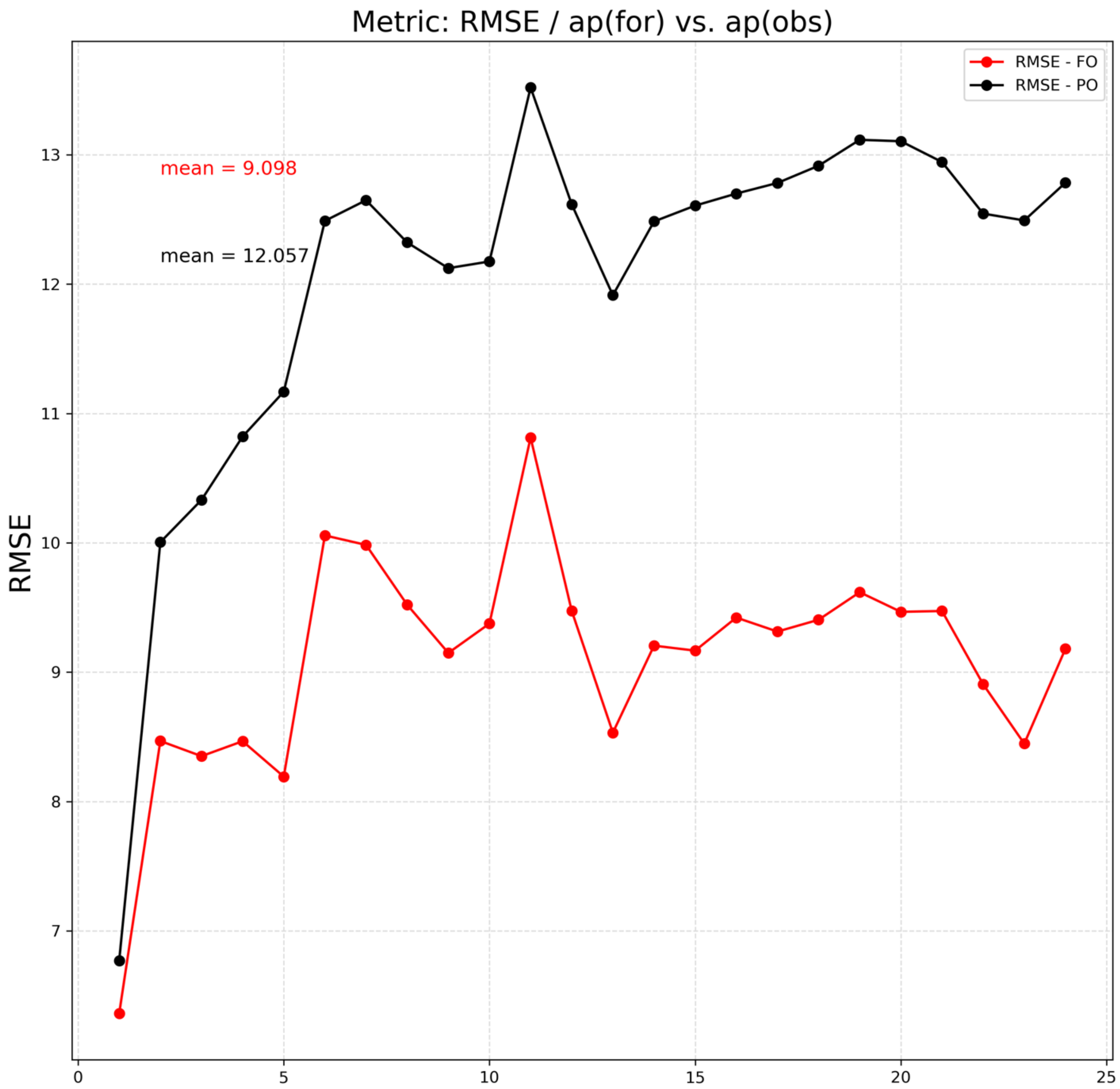
The model which has a MAE closer to 0.0 has a better performance. The superior performance of LSTM model is presented in the next figure. It is very encouraging the fact that our model is better than PERS model even for the first forecasted ap value. After that the better performance of LSTM is obvious. The mean MAE values for all the predicted values (24-time intervals) for LSTM and PERS models are 5.232 and 6.449 respectively.



MAE as a function of time intervals for LSTM model (red) and the PERS model (black). The better performance of LSTM vs. PERS model is obvious even for the first predicted value.

RMSE

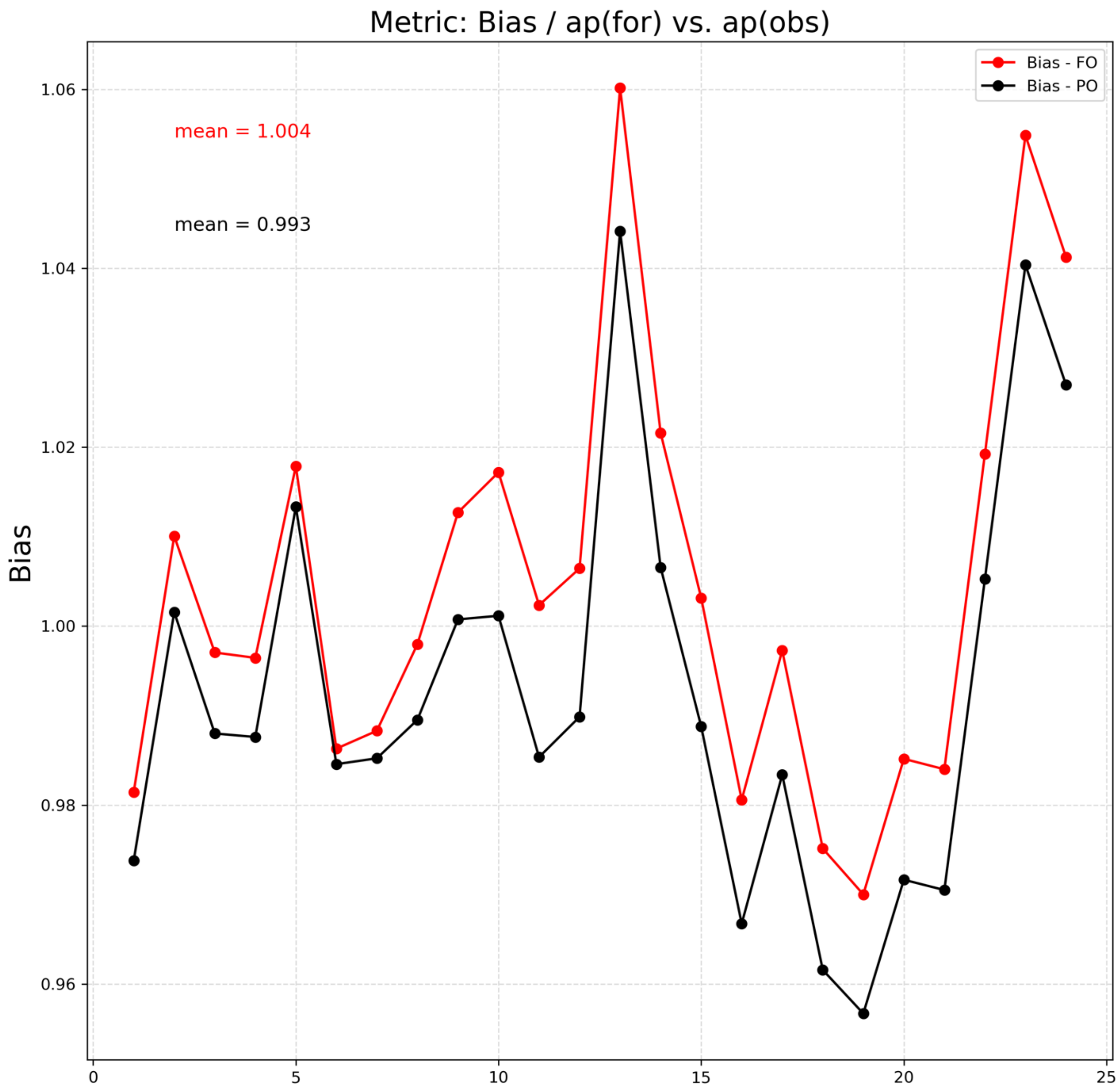
The best performance is for a RMSE value as closer as possible to 0.0. As for MAE the LSTM model has a better performance than PERS model (see next figure). The mean values of RMSE for all time intervals for LSTM and PERS models are 9.098 and 12.057 respectively.



RMSE as a function of time intervals for LSTM model (red) and the PERS model (black). The better performance of LSTM vs. PERS model is obvious even for the first predicted value as observed also in MAE and ME.

Bias

The model with a bias closer to 1 has a better performance. In this case we get similar results to the ME analysis. LSTM model has a better performance than the PERS model. The mean bias values for LSTM and PERS models are 1.004 and 0.993 respectively and in 62.5% of the examined time intervals the LSTM is closer to 1 than the PERS model.



Bias as a function of time intervals for LSTM model (red) and the PERS model (black).

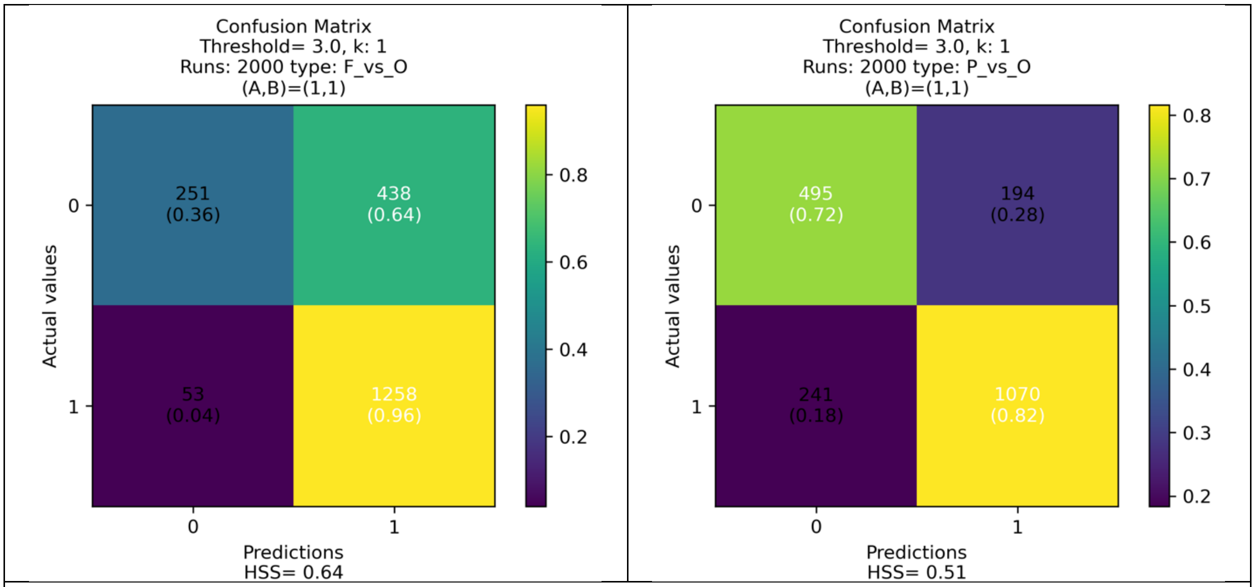
Concluding, in all metrics of Category I: Fit Performance Metrics the LSTM has a better performance in comparison to PERS model with an MAE of 5.2 (PERS 6.5) and an RMSE of 9.1 (PERS 12.1).

B. Threshold Performance Metrics

The ap index is not a continuous variable but is not a binary variable either, and this fact is increasing the complexity of the validation analysis. For a binary variable (e.g. “yes” vs. “no” result) is very common to calculate a series of metrics based on the contingency table. For continuous variable, it is more complex to create a binary “yes/no” situation. As a result, an index value serves as a “threshold” value (see e.g. Liemohn et al., 2018) to create this “yes/no” criterion. For the validation analysis we are using each value of ap index starting from 0 up to the very high and rare values of ap as a threshold value. With the known threshold ap value a contingency table could be created as follows:

- The term A is the number of the correct forecasted events or hits. When the predicted and observed values are less than or equal to the threshold value, this pair is considered as a “hit”.
- The term B is the number of false alarms. A false alarm is a forecast of an event, while no event was observed. When a forecasted value is less or equal with the threshold value, while the observed value was higher than the threshold, this pair is considered as a “false alarm”.
- The term C is the number of misses. A miss is an event that was not forecasted. In that case, the predicted value is higher than the threshold value, and the observed value is less or equal to the threshold value. This pair is considered as a “miss”.
- The term D is the number of true negatives or correct rejections. A correct rejection is a forecast of a non-event, while indeed, no event was observed. In that case, the forecasted and the observed values are greater than the threshold, and this pair is considered as “correct negatives”.

The sum of A+B+C+D is equal to the total number of pairs or the sample size (n). Two examples of contingency tables for LSTM and PERS model outputs are presenting below. In this case the output from 2000 runs are used to create these tables for a threshold ap value of 3 and for the first predicted ap value. This procedure performed for 17 threshold ap values [0, 56] and for each one of the forecast time windows starting from the first 3-hour time interval up to 72-hours (24th time interval) from T0. The top row of the table are the A (left) and C (right) terms and at the second row of the table are the B (left) and D (right) terms as described before. In each quartile we see the absolute numbers and the ratio for each case.

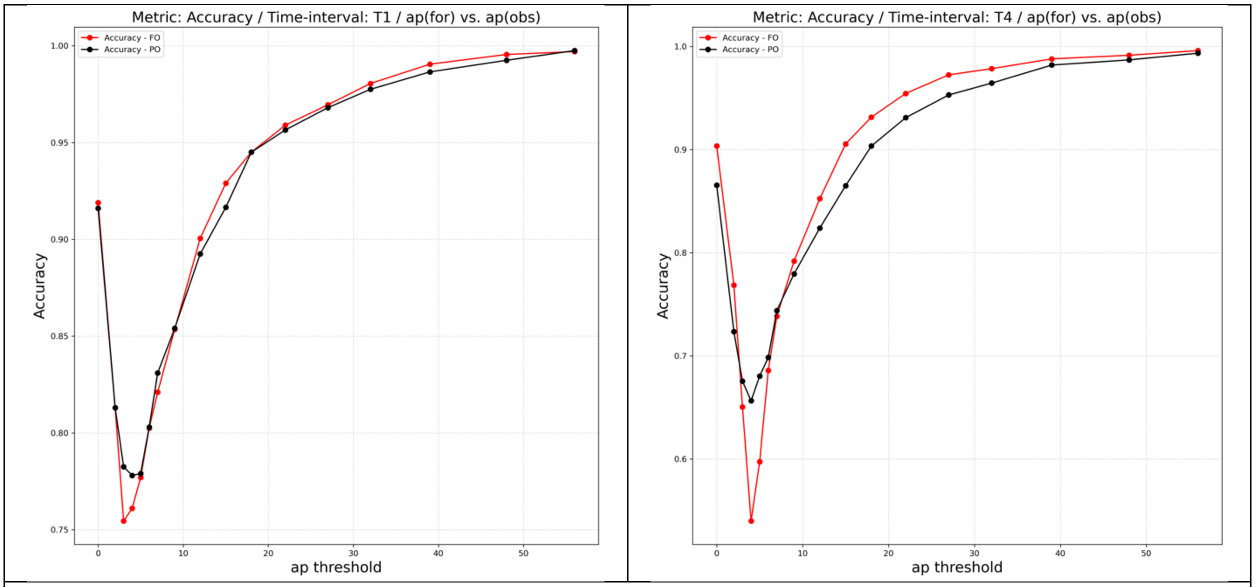


Contingency tables for a total of 2000 runs for LSTM (left) and PERS (right) models for threshold ap value 3 and for the first predicted time interval.

Various quantities and metrics are calculated using a contingency table. According to previous works (Jolliffe and Stephenson, 2012; Devos et al., 2014; Liemohn et al., 2018, Paouris et al., 2021) as well as metrics from WMO (at https://www.cawcr.gov.au/projects/verification/), a set of useful quantities and skill scores, concerning the validation between forecasted and observed values, is created. The threshold validation metrics which are calculated are: the accuracy, the probability of detection (POD), the false alarm ratio (FAR), the probability of false detection (POFD), the success ratio (SR), the threat score (TS), the Gilbert skill score (ETS), the Hansen and Kuipers discriminant (TSS) and the Heidke skill score (HSS). As stated before, there is no a standard set of verification metrics. We followed previous works on similar verification analysis and we have selected common metrics (see e.g. Devos et al., 2014; Liemohn et al., 2018 and references there in). Information on these metrics and their mathematical equations are presented in detail in Paouris et al., 2021.

Accuracy

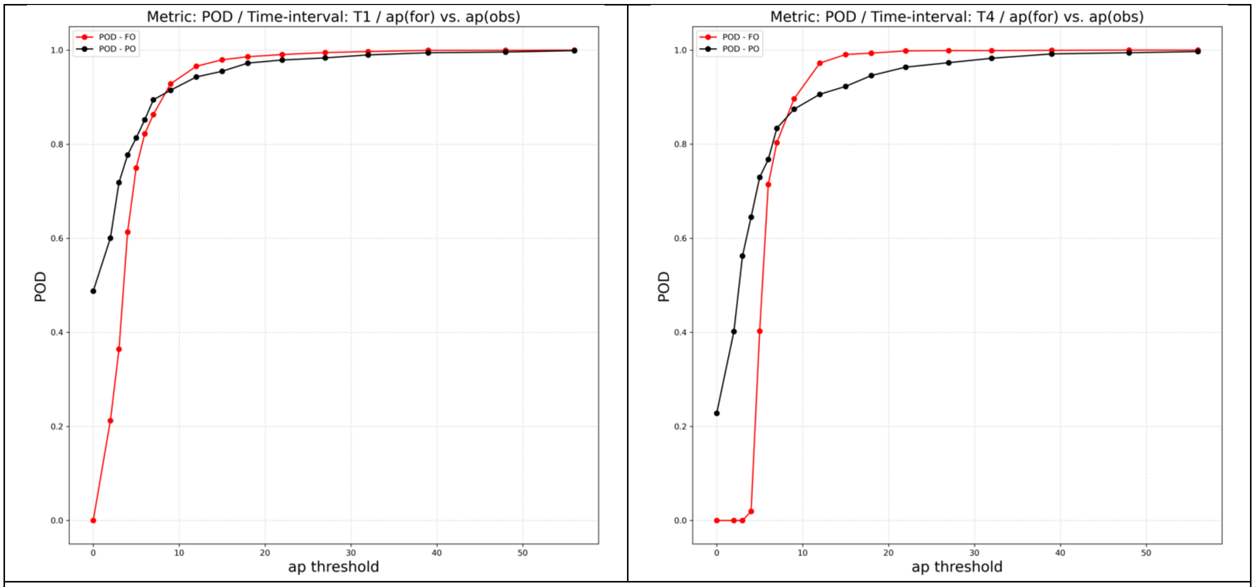
The closer the value of accuracy to 1 the better the model performs. The LSTM has a better performance than the PERS for all the threshold values above 9 for the first predicted value. This behavior is constant for all the predicted values. As an example, two plots of accuracy for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure. The PERS model has slightly higher accuracy values only for the very small values of ap [0,3] as expected for the very quiet periods.



Accuracy as a function of ap threshold levels for LSTM (red) and PERS (black) models.

POD

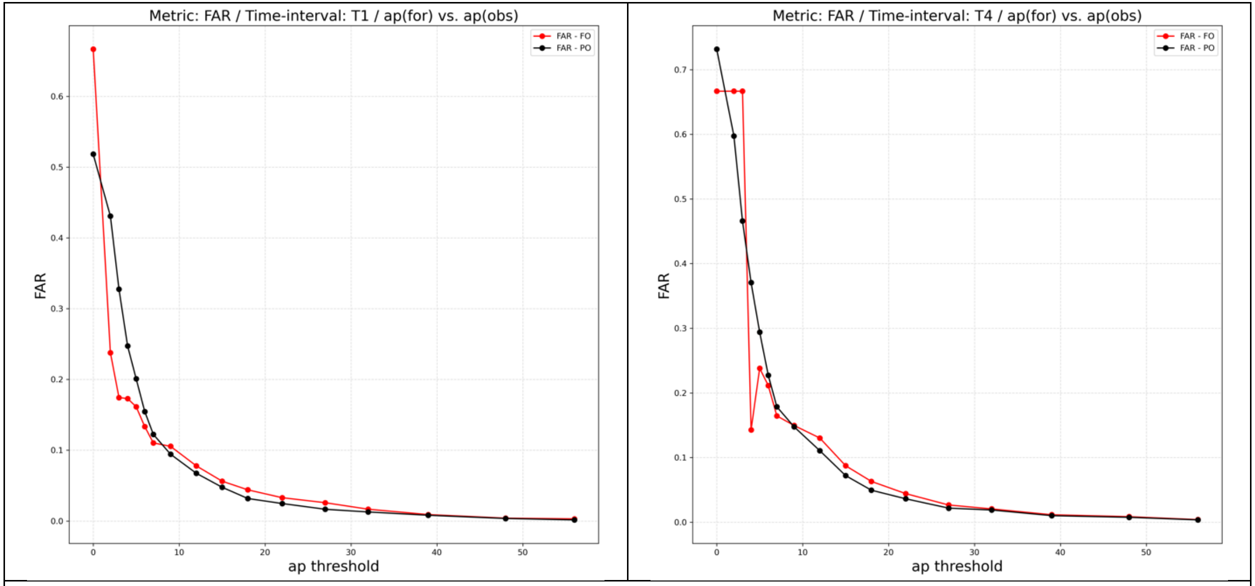
For the POD the value should be closer to 1 for best performance. In this case we get similar results to the accuracy. The LSTM model has almost perfect score above the threshold ap = 7. As an example, two plots of POD for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure.



POD as a function of ap threshold levels for LSTM (red) and PERS (black) models.

FAR

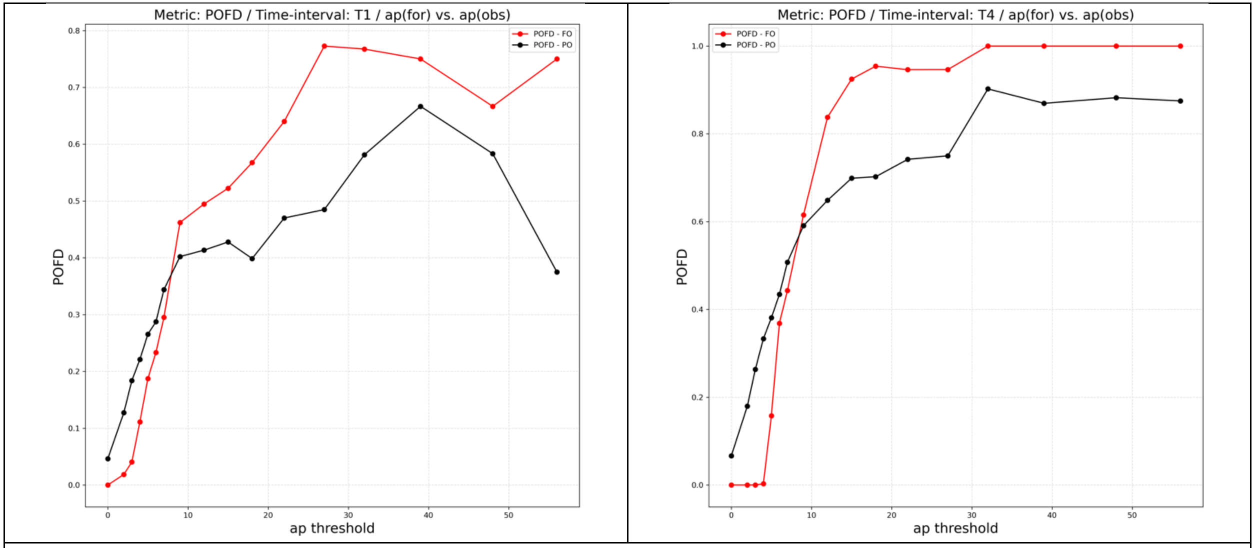
The value of 0 represents the best performance of the model. In this case we observe an opposite behavior for LSTM and PERS models. For lower values of ap (threshold < 9) the LSTM has a better performance than PERS with significant differences from PERS. Above that limit the PERS is above the LSTM but with similar numbers for FAR. As an example, two plots of FAR for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure.



FAR as a function of ap threshold levels for LSTM (red) and PERS (black) models.

POFD

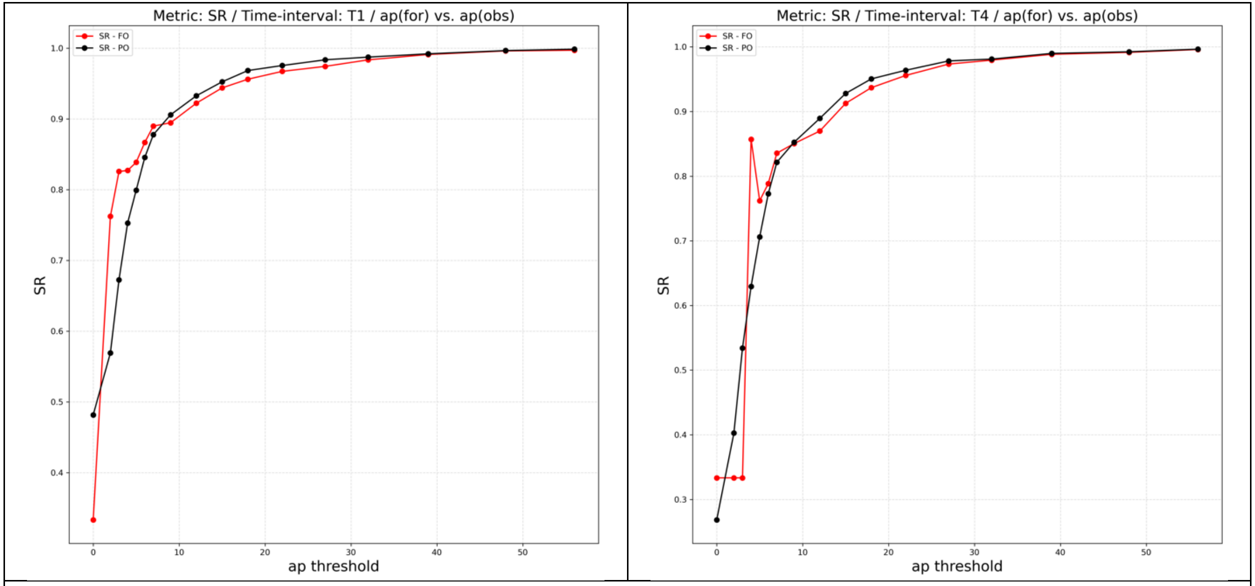
The best performance is for POFD = 0. Similar results to FAR are observed also for POFD. LSTM takes values closer to 0 for lower ap values (threshold < 9). As an example, two plots of POFD for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure.



POFD as a function of ap threshold levels for LSTM (red) and PERS (black) models.

SR

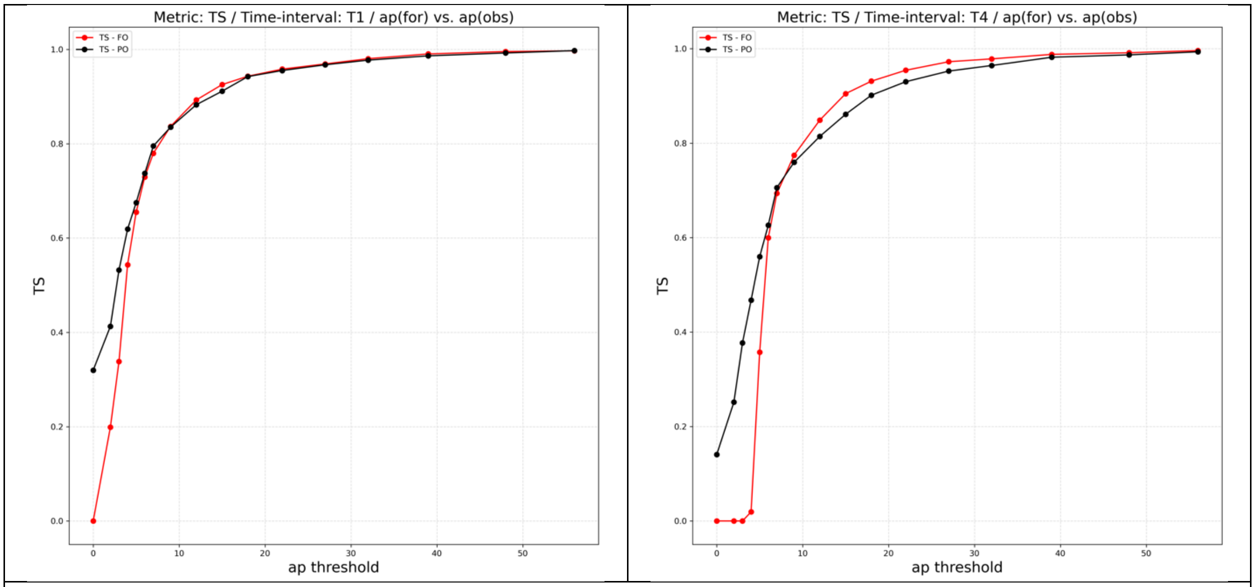
The best performance for a given model is SR = 1. In this case the results are similar to the accuracy. LSTM model has a better performance from the PERS model for ap threshold values less than 9. Above this threshold PERS has slightly greater SR values. As an example, two plots of SR for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure.



SR as a function of ap threshold levels for LSTM (red) and PERS (black) models.

TS

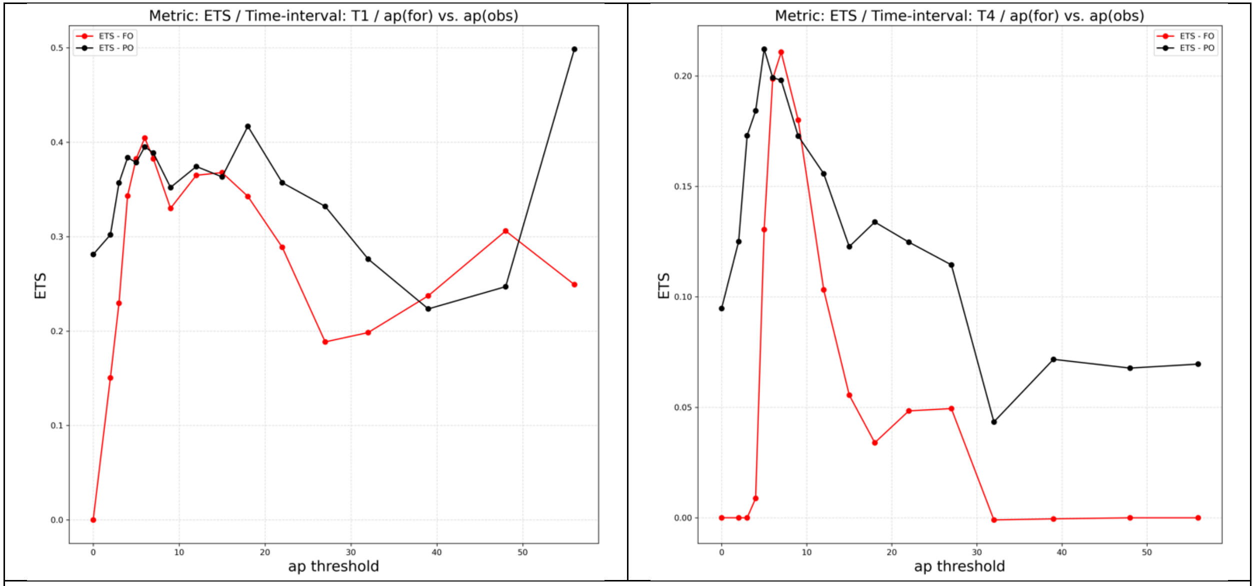
The best performance is for TS = 1. For this metric and particularly for the first ap prediction (T1) both models has similar results (LSTM is slightly better) above the threshold ap > 9 while the PERS has better output for small ap values. As the time interval increases (T2, T3, T4,…) the LSTM has a better performance in comparison to the predicted ap for T1 for the threshold > 9. As an example, two plots of TS for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure.



TS as a function of ap threshold levels for LSTM (red) and PERS (black) models.

ETS

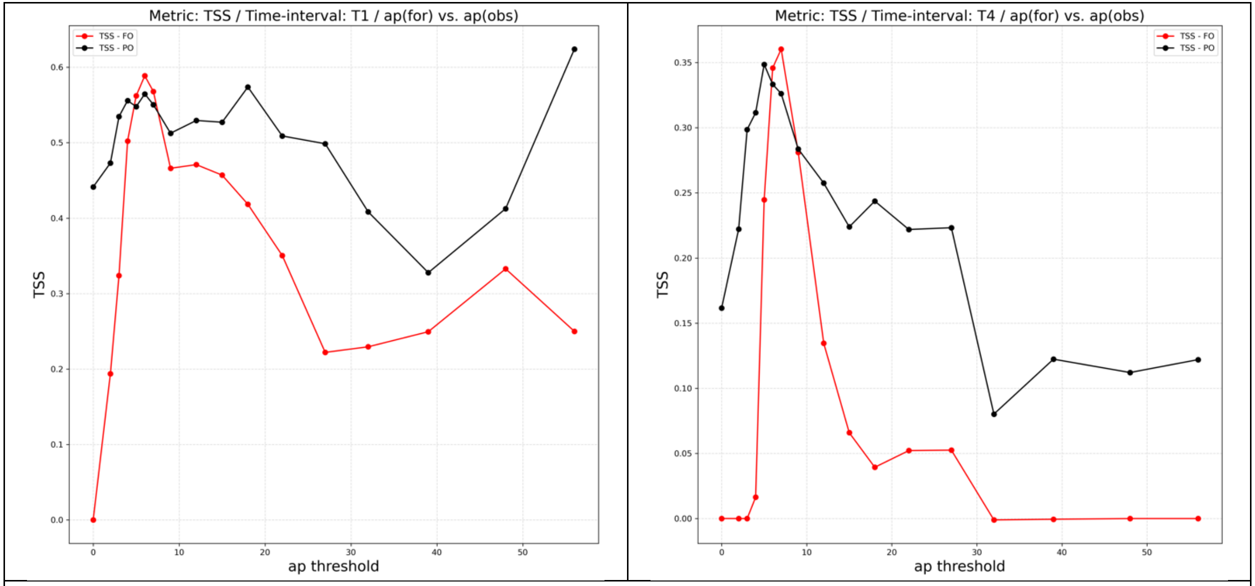
Best performance for ETS values equal to 1. For this metric (as well for other similar metrics) we expect different values for different thresholds (see e.g. Paouris et al., 2021). The output of T1 does not present a clear trend. LSTM has the maximum value of 0.41 with threshold ap = 6 and this value is the highest in contrast to PERS model. The maximum ETS = 0.42 is observed for the threshold ap = 18 for PERS model. The 0.5 for PERS for threshold ap = 56 is an outlier and is associated with two misses (term C) which observed in the sample of 2000 runs. As the forecasts windows increasing the ETS is taking lower values. LSTM and PERS models has similar maximum ETS value of ~0.22. As an example, two plots of ETS for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure.



ETS as a function of ap threshold levels for LSTM (red) and PERS (black) models.

TSS

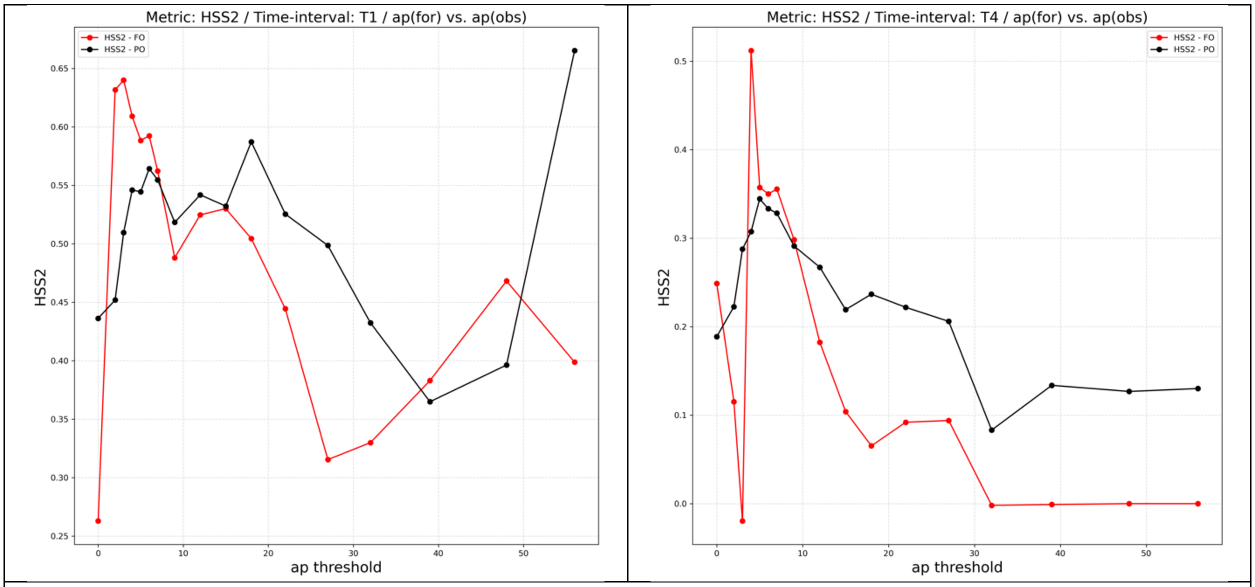
The best performance for the Hansen and Kuipers discriminant is represented for a value equal to 1. The best performance is observed for all cases for LSTM, except the same outlier we mentioned before. Even for the outlier the difference between LSTM and PERS is very small (0.59 vs. 0. 62). The trend for LSTM and PERS is very similar with decrease as the ap threshold increasing and the TSS values are decreasing as the forecast window is increasing. As an example, two plots of TSS for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure.



TSS as a function of ap threshold levels for LSTM (red) and PERS (black) models.

HSS

As before, the best performance model is for Heidke skill score values closer to 1. LSTM has a better performance for ap threshold values less than 9 and the PERS for ap greater than 9. The overall maximum HSS for LSTM is 0.64 while for PERS is 0.58 (except the outlier) for the first predicted ap (T1). The trend is similar to the previous metric (TSS). For T2 the maximum values are 0.58 (LSTM) and 0.48 (PERS), for T3 0.52 (LSTM) and 0.45 (PERS) and finally for T4 are 0.52 (LSTM) and 0.34 (PERS). As an example, two plots of HSS for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure



HSS as a function of ap threshold levels for LSTM (red) and PERS (black) models.

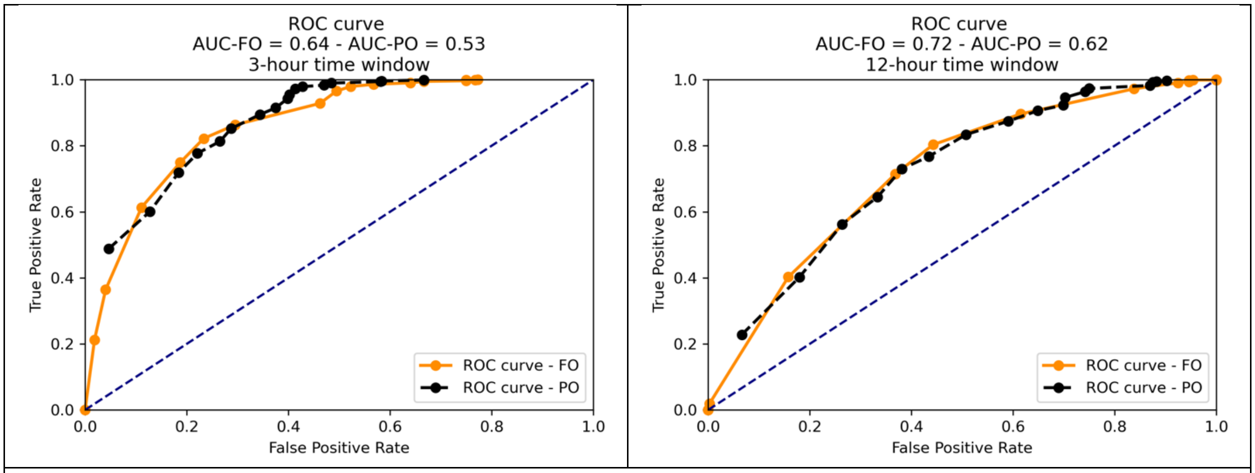
ROC

An essential element in the analysis of the event detection assessment, is the receiver operating characteristic (ROC) curve that shows the ability of the forecast to discriminate between events and non-events. The ROC curve plot was created as follows: the probability of detection (POD) set on the y-axis and the probability of false detection (POFD) placed on the x-axis for all the threshold values of ap geomagnetic index. The ideal model curve should travel from bottom left to top left of the diagram and then across to the top right of the diagram. As the closer the curve is to the unity slope line represents no skill. The area under the curve for a perfect score equals 1.

In this case the LSTM model shows better performance than PERS in almost every case for the first 8 predictions (T1 – T8). After the first 24 hours both models have similar AUC values around 0.55 indicating not very well predictions skills. This is not a surprise as in most cases of published machine learning models the forecasting window is ranging from a few minutes up to a few (1-6) hours (see e.g. Camporeale, 2019 and references therein and the most recent work Hu et al., 2022). In order to make it simpler we provide also the are under the curve (AUC) for each time forecast window in the next table:

Forecast time window	**AUC / LSTM**	**AUC / PERS**
3-hour time interval T1	0.64	0.53
6-hour time interval T2	0.73	0.65
9-hour time interval T3	0.77	0.64
12-hour time interval T4	0.72	0.62

As an example, two plots of ROC and AUC for the first (T1) and the fourth (T4) predicted values of ap as a function of threshold ap values are presented in the next figure



ROC and the calculated AUC for the first and the fourth predicted ap values (T1 and T4) for the LSTM (orange) and the PERS (black) models. The dichotomous y=x (blue dashed line) indicates no skill for prediction.

Concluding, the analysis of the metrics of Category II: Threshold Performance Metrics shows that the LSTM has a better performance in comparison to PERS model in almost every metric particularly with a maximum value of HSS of 0.64 (PERS 0.58) and a maximum AUC value of 0.77 (PERS 0.65). The Category II metrics are more complicated than the ones of Category I and should be used as an important supplement for Category I.

The results of this validation analysis will be submitted in the next few weeks for publication in a refereed journal.

References

Bailey, R.L. *et al.*, “Forecasting GICs and Geoelectric Fields from Solar Wind Data Using LSTMs: Application in Austria,” *Space Weather*, vol. 20, no. 3, Mar. 2022, doi: 10.1029/2021SW002907.

Camporeale, E., “The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting,” *Space Weather*, vol. 17, no. 8, pp. 1166–1207, Aug. 2019, doi: 10.1029/2018SW002061.

Devos, A., Verbeeck, C., Robbrecht, E., 2014. Verification of space weather forecasting at the regional warning center in Belgium. Space Weather Space Clim 4 (27). https://doi.org/10.1051/swsc/2014025. A29.

Jolliffe, I.T., Stephenson, D.B., 2012. Forecast verification. A practitioners Guide in Atmospheric Science, 2nd ed. Wiley-Blackwell, p. 2012.

Hu, A.,Camporeale, E., & Swiger, B. (2023). Multi-hour-ahead Dst index prediction using multi-fidelity boosted neural networks. *Space Weather*, 21, e2022SW003286. https://doi.org/10.1029/2022SW003286

Liemohn, M.W., McCollough, J.P., Jordanova, V.K., Ngwira, C.M., Morley, S.K., Cid, C., et al., 2018. Model evaluation guidelines for geomagnetic index predictions. Space Weather 16, 2079–2102. https://doi.org/10.1029/2018SW002067.

Paouris, E., M. Abunina, A. Belov, and H. Mavromichalaki, “Statistical analysis on the current capability to predict the Ap Geomagnetic Index,” *New Astronomy*, vol. 86, p. 101570, Jul. 2021, doi: 10.1016/j.newast.2021.101570.