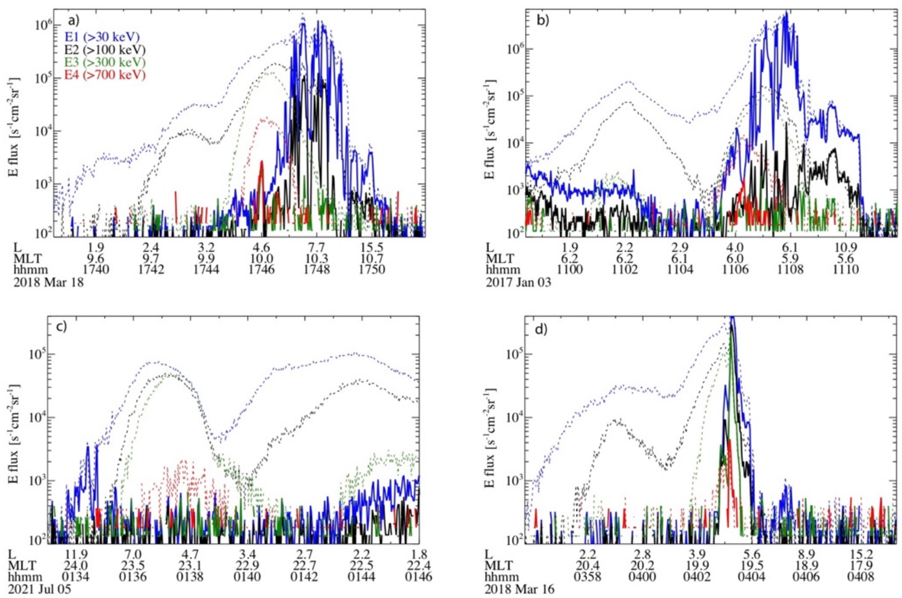
Supplementary Material

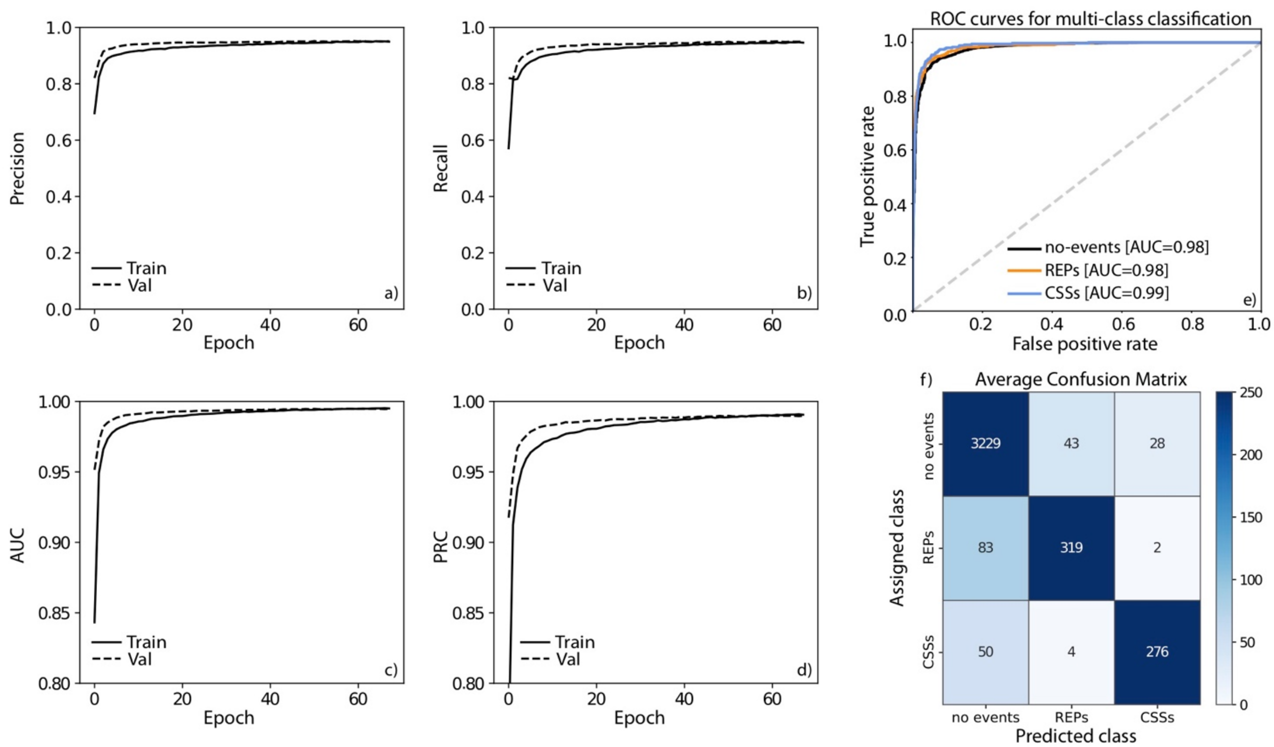
1 Supplementary Figures and Tables

In this section, we illustrate three figures. Supplementary Figure 1 shows the variability of POES data over ¼ orbit, for four different days. Supplementary Figure 2 shows the metrics relative to one k-fold and the confusion matrix resulting from the k-fold cross validation. Supplementary Figure 3 shows precipitation events identified by the model that might require additional post-processing to be handled.

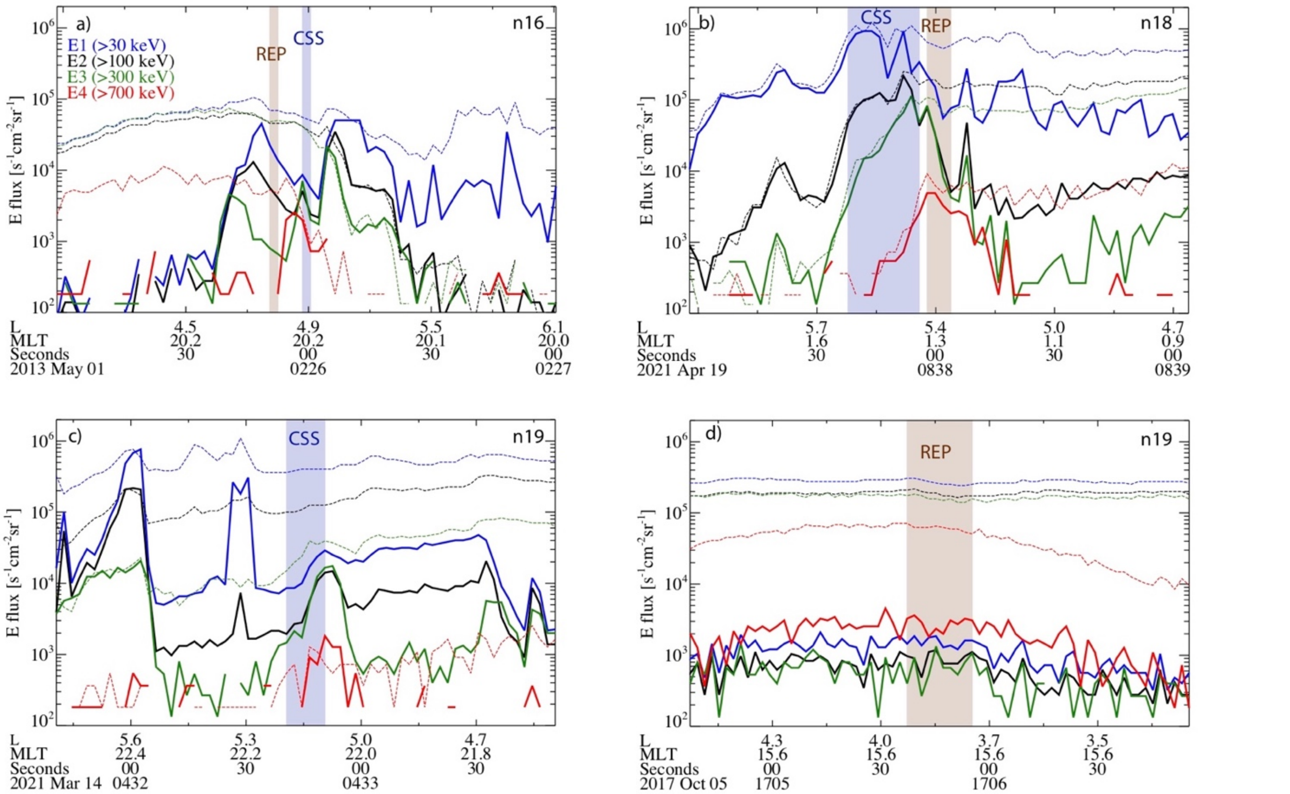
1.1 Supplementary Figures



Supplementary Figure 1. POES/MetOp data over ¼ orbit for four different days: a) 18 March 2018 from MetOp01 (m01), b) 3 January 2017 from NOAA-18 (n18), c) 5 July 2021 from NOAA-18 (n18), and d) 16 March 2018 from MetOp-01 (m01). Dashed lines indicate the 90° telescope measurement (trapped electrons) and the solid lines indicate the 0° telescope measurements (precipitating electrons). Electron channels are shown for different energies as indicated by the legend in panel a). These figures show that over ¼ orbit POES/MetOp observe a high variability and complexity of electron fluxes, both trapped and precipitating. Additionally, since the REP or CSS events are rare and short-lived (for example, REP at 17:46 UT in a) and CSS at 04:03 UT in d)), including a full day of POES/MetOp data for each REP or CSS event would cause significant class imbalance (towards the “no-event” class). Due to the POES orbital period (~100 min), there are ~60 quarter orbits per day with patterns similar to those shown here, in which each data point would have to be classified as a “no-event” (except if an event is found as in a) or d)). This is one of the reasons why our dataset only includes a limited number of “no-event” data points adjacent to the REP or CSS.



Supplementary Figure 2. Panels a)–e) represent the metrics for the third k-fold training. In the main paper, we mention we obtained the model performances using a k-fold cross validation (k=10) and here we show an example of the precision (a), recall (b), AUC (c), AUPRC (d) and ROC for each class (e) for one k-fold (k=3). Panel f) shows the confusion matrix resulting from averaging all the confusion matrices of each k-fold. The confusion matrix shows a comparison between the manually assigned classes (actual class, y axis) and the class predicted by the model (predicted class, x axis). The color bar indicates how many data snapshots belong in each square, which are written in each box. A confusion matrix shows high values along its diagonal if the classification is correct. In this case, the confusion matrix shows the highest values along the diagonal, indicating that the highest numbers of predicted no-events, REPs and CSSs indeed belong to their respective class.



Supplementary Figure 3. In a similar format to Figure 4 in the main paper, we show the events identified and classified by the model for four days of POES/MetOp data. Panels a) and b) show that the model sometimes identifies one event in near proximity of another one, each belonging to different classes. The events in panel a) are both short-lived and probably the correct class of the precipitation event here is a REP because there is not coincident energy-dispersion at lower electron energy. Nevertheless, there is some energy dispersion at ~02:25:45 UT (likely due to proton contamination) and some energy-dependent precipitation after 02:26 UT that might have triggered the model to classify this event as both a REP and a CSS. The scenario is less complex in panel b) where this event is a true CSS. The model is still indicating that part of the precipitation is a REP, which is not correct in this case. Panel c) and d) are also likely false positives indicated by the model. The event in panel c) does not show clear energy-dependent precipitation, thus it is mistakenly classified as a CSS, while it should be a REP. The event in panel d) is not showing the isolated precipitation we expect from a REP, thus it is a false positive. We suspect the data in this POES/MetOp time interval is not reliable because the precipitating relativistic electron flux is higher than the lower energy electron fluxes, thus it would have to be ruled out. The purpose of this figure is to demonstrate that no matter how high the performance score of a model is, a model always has limitations. Post-processing of the model outputs is needed before being able to use them for scientifical statistical studies on understanding the distribution of the REP *vs.* CSS precipitation. In this way, we could limit the number of false positives or mistakenly classified events. This is beyond the scope of the current paper and is left as a future task.

1.2 Supplementary Table

Table 1. This table illustrates the performance scores (rows: F1, AUC, AUPRC) for each of the architectures tested (columns) in this work. The scores are reported after the k-fold cross-validation with k=10. The 64 bdir-LSTM + 256 dense architecture has the highest performance scores and it was selected for our study.

Architecture /Scores	**64 bdir-LSTM + 256 dense**	**64 bdir-LSTM + 64 bdir-LSTM + 128 dense**	**128 LSTM + 128 dense**	**128 LSTM + 256 dense**	**64 LSTM + 256 dense**
F1	0.9480	0.9467	0.9461	0.9455	0.9432
AUC	0.9947	0.9946	0.9944	0.9944	0.9941
AUPRC	0.9898	0.9897	0.9893	0.9892	0.9888