An operational approach to forecast the Earth's radiation belts dynamics

Guillerme Bernoux, Antoine Brunet, Éric Buchlin, Miho Janvier, Angélica Sicard

submitted to **Journal of Space Weather and Space Climate**

An operational approach to forecast the Earth's radiation belts dynamics

Guillerme Bernoux[1,⋆], Antoine Brunet[1], Éric Buchlin[2], Miho Janvier[2], and Angélica Sicard[1]

[1] ONERA / DPHY, Université de Toulouse, F-31055 Toulouse, France

[2] Université Paris-Saclay, CNRS, Institut d'Astrophysique Spatiale, Orsay, France

ABSTRACT

The *Ca* index is a time-integrated geomagnetic index that correlates well with the dynamics of high-energy electron fluxes in the outer radiation belts. Therefore *Ca* can be used as an indicator for the state of filling of the radiation belts for those electrons. *Ca* also has the advantage of being a ground-based measurement with extensive historical records. In this work, we propose a data-driven model to forecast *Ca* up to 24 hours in advance from near-Earth solar wind parameters. Our model relies mainly on a recurrent neural network architecture called Long Short Term Memory that has shown good performances in forecasting other geomagnetic indices in previous papers. Most implementation choices in this study were arbitrated from the point of view of a space system operator, including the data selection and split, the definition of a binary classification threshold, and the evaluation methodology. We evaluate our model (against a linear baseline) using both classical and novel (in the space weather field) measures. In particular, we use the Temporal Distortion Mix (TDM) to assess the propensity of two time series to exhibit time lags. We also evaluate the ability of our model to detect storm onsets during quiet periods. It is shown that our model has high overall accuracy, with evaluation measures deteriorating in a smooth and slow trend over time. However, using the TDM and binary classification forecast evaluation metrics, we show that the forecasts lose some of their usefulness in an operational context even for time horizons shorter than 6 hours. This behaviour was not observable when evaluating the model only with metrics such as the root-mean-square error or the Pearson linear correlation. Considering the physics of the problem, this result is not surprising and suggests that the use of more spatially remote data (such as solar imaging) could improve space weather forecasts.

Key words. space weather – forecasting – radiation belts – machine learning – solar wind

1. Introduction

One of the current main topics of interest in the space weather field is the forecasting of geomagnetic indices based on machine learning methods. Machine learning has allowed for a great improvement

⋆ Corresponding author: guillerme.bernoux@onera.fr

This provisional PDF is the accepted version. The article should be cited as: J. Space Weather Space Clim., doi: 10.1051/swsc/2021045

in short-term forecasts of geomagnetic indices such as the global index *Kp* (Wintoft et al., 2017; Tan et al., 2018; Chakraborty and Morley, 2020) or *Dst* index (Gruet et al., 2018; Lethy et al., 2018). Space weather-induced events can have heavy-to-extreme consequences on human-made infrastructures, as for instance space-borne hardware or even ground-based facilities (Riley et al., 2017). That is why the reliable forecast of geomagnetic indices and other space-weather relevant physical quantities (e.g. relativistic electron or proton fluxes in the radiation belts) is of paramount importance.

The extent of the effects of the space radiative environment on satellites ranges from single events caused by high energy charged particles from cosmic rays or solar energetic particles (SEP) to internal charging, surface charging, or total ionising dose (Horne et al., 2013). Therefore, being able to accurately and reliably forecast the fluxes of high-energy electrons (from dozens of kiloelectronvolts to a few megaelectronvolts) in the radiation belts would represent a great leap towards better mitigation of the radiation-induced risks in space. Extensive efforts have already been conducted to forecast such electron fluxes. A considerable review of the methods used to forecast these electron fluxes was recently proposed by Camporeale (2019), where it is detailed that feed-forward neural networks and recurrent neural networks (RNNs) are used to obtain forecasts up to a few hours or a few days ahead (see e.g. Ling et al. (2010); Wei et al. (2018)).

However, Camporeale (2019) notes that although many approaches have been tested, it remains difficult to predict these fluxes due in particular to certain physical phenomena that are difficult to take into account for a "black-box" type model. Thus, many more recent models based on machine learning methods do not seem to perform better than older models. In addition, using data-driven approaches to predict radiation belt dynamics with in-situ data is challenging since it is important to have large databases that are properly calibrated (which is more complicated when using space-borne instruments rather than ground-based ones).

Recently, Bernoux and Maget (2020) have proposed a new time-integrated geomagnetic index that aims to be more representative of the state of filling of the Earth's radiation belts. This so-called *Ca* index is a time-integrated index based on the better-known *aa* index. As we will see in detail in section 2, *Ca* was created to take into account the intensification of trapped electrons in the radiation belts. *Ca* is therefore a complementary index to other indices such as *Kp* or *Dst*. Thus in this study we focus on the prediction of the radiation belts dynamics represented by the *Ca* index. To do so we will use deep learning methods (i.e. machine learning approaches based on deep neural networks) that have already been successfully tested with other geomagnetic indices. However, and in contrast to other studies, we concentrate on evaluating our models by trying to take into account the point of view of a spacecraft operator. Therefore we use evaluation methods other than the classical metrics such as the root-mean-square error and the linear correlation, which can only account for global behaviour and are consequently largely insufficient to quantify other phenomena such as time shifts.

In this work, we design a neural network-based model to forecast the *Ca* index up to 24 hours in advance. Then we evaluate the model using both classical metrics and also a method to detect the systematic existence of time shifts in our predictions. We also transform the regression problem into a binary classification problem aimed at predicting danger periods in terms of surface charging and we evaluate it accordingly. In Section 2 we present the data sets used in our models and we explain why they were chosen and how they were pre-processed. In Section 3 we present the models

and their dedicated evaluation methods. In Section 4 we present and discuss the results before concluding in Section 5.

2. Data analysis

In this section, we describe and analyse the data sets used in this paper. Firstly we list the solar wind parameters and geomagnetic indices used here and explain where and how they can be obtained. Then we focus on the geomagnetic index *Ca* and explain its relevance to our purposes. Finally, we explain how the time periods used for the training and the evaluation of the different models were selected.

2.1. Data sets

It is now well known that the geomagnetic indices representing the state of the magnetosphere are predominantly driven by solar wind dynamics (Akasofu, 1981; Baker et al., 1981). That is why, as in many other studies (e.g. Lundstedt and Wintoft, 1994; Wu and Lundstedt, 1997; Wing et al., 2005; Chandorkar et al., 2017; Chakraborty and Morley, 2020), we use solar wind parameters available in the OMNIweb database (King and Papitashvili, 2005) as inputs to our geomagnetic index forecast models. The OMNIweb database (`https://omniweb.gsfc.nasa.gov/`) grants access to hourly spacecraft-interspersed near-Earth measurements of solar wind parameters. Earliest solar wind parameters are available since late 1963. In particular we select the plasma bulk velocity V_{sw}, the ion density ρ, the southward component of the interplanetary magnetic field (IMF) B_z and the plasma temperature T as the inputs to our models. It is now well known that these parameters correlate well with geomagnetic indices and with the dynamics of electron fluxes in the radiation belts (Burton et al., 1975; Wing et al., 2016). A thorough study based on information-theoretical tools could help us in finding an even better set of input parameters, but this is out of the scope of our study and could be the topic of future work.

The geomagnetic index studied here is the *Ca* index, that was first introduced by Bernoux and Maget (2020) based on a previous study by Rochel et al. (2016). Therefore the following paragraphs rephrase some information on the purpose and relevance of this index that was contained in these papers.

The *Ca* index is an index derived from the well-known *aa* index. The *aa* index is a 3-hr K-based index first introduced by Mayaud (1971) and computed from data provided by two subauroral antipodal observatories. *aa* index is the geomagnetic index having the longest available track record with data available since 1868. This gives us more than 150 years of homogeneous (Mayaud, 1980) and exploitable geomagnetic data with a time cadence of 3 hours. This is particularly useful when dealing with topics, such as statistical analysis, which require a great amount of data. In particular, *aa* index covers a time range equivalent to 14 solar cycles. Nowadays, the *aa* index is made available by the International Service of Geomagnetic Indices (ISGI) and can be downloaded from their website (`http://isgi.unistra.fr/data_download.php`).

As stated in Bernoux and Maget (2020), *Ca* index has been designed to quantify the geoeffectiveness of solar wind structures impacting the magnetosphere from the radiation belts perspective. The relaxation characteristic time in the radiation belts for high-energy electrons after a strong magnetospheric disturbance is of the order of 4 days (Meredith et al., 2006; Rochel et al., 2016). Therefore

the *Ca* index is defined as follows:

$$Ca(t) = \frac{1}{\tau}\int_0^{\infty} aa(t-t')\mathrm{e}^{\frac{-t'}{\tau}}dt' \quad (1)$$

with $\tau = 4$ days being the relaxation characteristic time and *aa* representing the geomagnetic activity. Being directly derived from *aa* index, *Ca* index shares the same above-mentioned qualities and properties. Further details on the interest and relevance of using the *Ca* index are provided in subsection 2.2.

2.2. *Why study and forecast the* Ca *index?*

Numerous studies have already been conducted on the topic of the nowcasting and forecasting of geomagnetic indices. Many of them focus on the *Kp* index or the *Dst* index, which are two very well-known indices that have been thoroughly studied for decades. However, it should be reminded that all geomagnetic indices are not interchangeable and that those indices have physical meanings. For instance, Borovsky and Shprits (2017) makes clear that the *Dst* index is unable to capture all types of geomagnetic storms behaviour and is in reality a very poor index when studying space-weather-relevant phenomena such as the dynamics of the electrons in the outer radiation belts induced by long-duration Corotating Interaction Regions (CIR)-driven storms. This is why it is important not to direct the research effort solely to the problem of forecasting the *Kp* and *Dst* indices, but to diversify the indices studied, in order to include a greater diversity of space-weather-relevant phenomena.

The *Ca* index was created to account for geomagnetic storms during which intensification of relativistic electrons trapped in the radiation belts is observed. It was shown in Rochel et al. (2016) and Bernoux and Maget (2020) that this index correlates well with electron fluxes ($E > 30$ keV) in the radiation belts and is able to take into account phenomena such as energy accumulation due to long-duration Stream Interaction Region (SIR)-driven storms, but also due to multiple successive Interplanetary Coronal Mass Ejection (ICME)-driven events. Figure 1 displays examples of the typical behaviour of the *Ca* index during ICME- and SIR-driven storms. During ICME-driven storms, the *aa* index tends to reach higher values (in this example *aa* reaches 228 nT) quickly, but it also decreases rapidly, whereas during SIR-driven storms the disturbance lasts longer even though the *aa* index usually does not reach such high values (in this example it only reaches 81 nT). Therefore the *Ca* index reaches its peak value much faster during the ICME-driven storm. However, the value of the peak is similar during both these events as *Ca* accounts better for energy accumulation (48.6 nT during the ICME-driven storm against 42.4 nT during the SIR-driven storm).

It was also stated in those papers that by changing the value of the parameter τ it is possible to easily create an index that accounts better for a given specific orbit (but then less for the others). It is interesting to note that the *Ca* index is not the only attempt to create an index with such properties and another approach was proposed by Borovsky and Yakymenko (2017).

From an operational perspective, the prediction of the *Ca* index could serve as a basis for an alert service for the accumulation of high-energy electrons in the radiation belts. In such a context, the *Ca* index would act as a proxy for relativistic electron fluxes, which is monitored from ground-based magnetometers. As stated in Section 1, using a data set that already has decades of cross-calibrated samples is also a great asset when dealing with data-driven approaches that require lots of data to be efficient. Besides, it may also be more reliable in terms of continuity of service to rely on ground-based instruments rather than onboard instruments that are subject to the risks associated with their

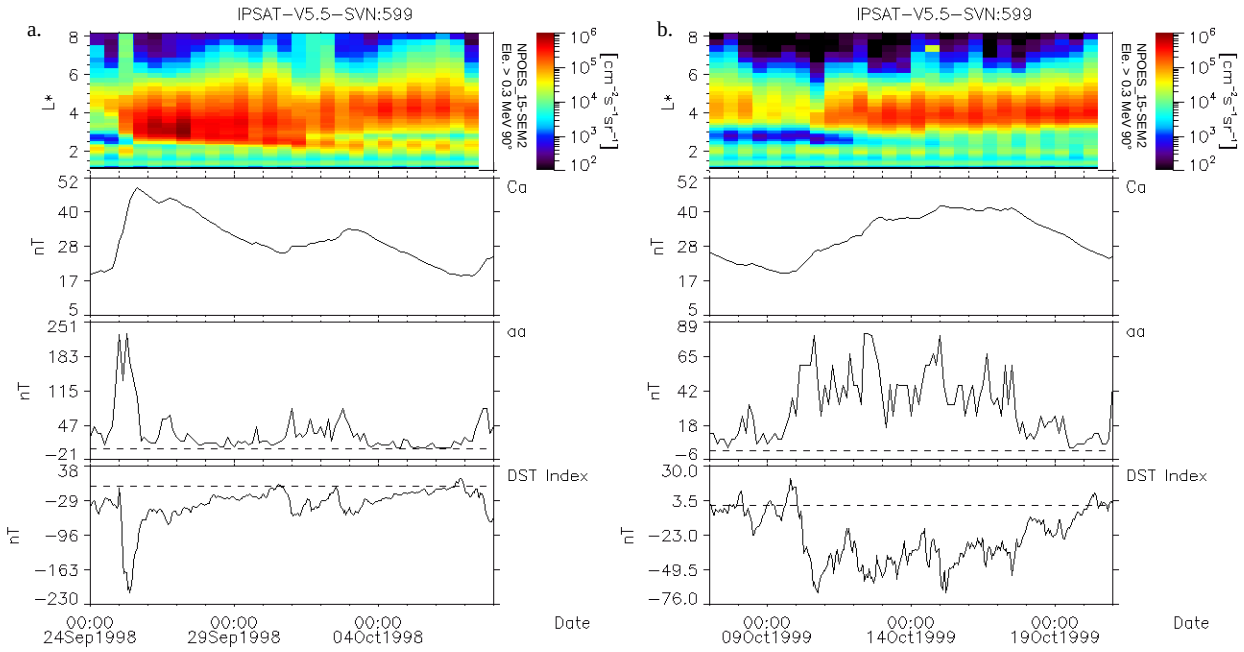

Fig. 1. From bottom to top: evolution of the geomagnetic indices *Dst*, *aa*, and *Ca*, and of the flux of electrons in the radiation belts for the $E \geq 300$ keV energy range, measured by the SEM instrument aboard the POES-15 spacecraft a) from 24 September to 8 October 1998 during a period that displayed an ICME-induced disturbance starting on 25 September 1998, and b) from 7 October to 21 October 1999 during a period that displayed a SIR-induced disturbance starting on 9 October 1999.

being in space, at least as a backup. Thus, the prediction of the *Ca* index is of immediate interest to the operators of space-borne systems.

2.3. Establishing the training, validation, and test sets

2.3.1. Splitting the data sets

In this subsection, we briefly analyse the time series supplied by the OMNIweb database in order to detect any important data gaps (that would be prejudicial for the training of a machine learning algorithm) and to carefully choose the time periods used to train, validate and evaluate our models. Dividing a data set into training, validation, and test sets is a very common practice in machine learning applications. If needed the reader is referred to Carè and Camporeale (2018) for more details.

Before the availability of the Wind/Solar Wind Experiment (Wind/SWE) and the Advanced Composition Explorer magnetometer and Solar Wind Electron, Proton, and Alpha Monitor (ACE/MAG and ACE/SWEPAM) data starting in 1995 and 1998, the OMNIweb database has a high percentage of missing data. Therefore in our study, we only use data from 1995 onward. For the 1995-2019 period there was on average 2.41% of missing data per year. Even if most of the gaps are very short ones, there are some gaps larger than three or four days, which require proper handling. That is why we decided to fill the data gaps with the method introduced in Kondrashov

et al. (2010). This method is based on Singular Spectrum Analysis, a data-adaptive spectral estimation method that is designed to provide information on the underlying dynamics of a (multivariate) time series (Ghil et al., 2002). In the context of space physics, SSA has already been used to fill the gaps in the OMNIweb database, which improved the accuracy of empirical magnetic field models compared to another simpler method based on linear interpolations (Kondrashov et al., 2014). Appendix A provides more information on the practical gap-filling of the time series used in this paper with a dedicated toolkit (Vautard et al., 1992).

The choice of the data that is used to train, validate and test the neural network is of critical importance. This includes the appropriate choice of how the data set is temporally subdivided into training, validation, and test data sets (Lazzús et al., 2017). In order to correctly train a machine learning algorithm, the training data set should be comprised of a representative period during which all kinds of space weather phenomena, including extreme events, were observed. The testing (and the validation) period should also be comprised of both quiet and agitated periods. Eventually, we have chosen the following periods, highlighted in Figure 2:

- Training set: 2003-01-01 – 2018-12-31
- Validation set: 1995-01-01 – 1996-12-31
- Test set : 1997-01-01 – 2002-12-31

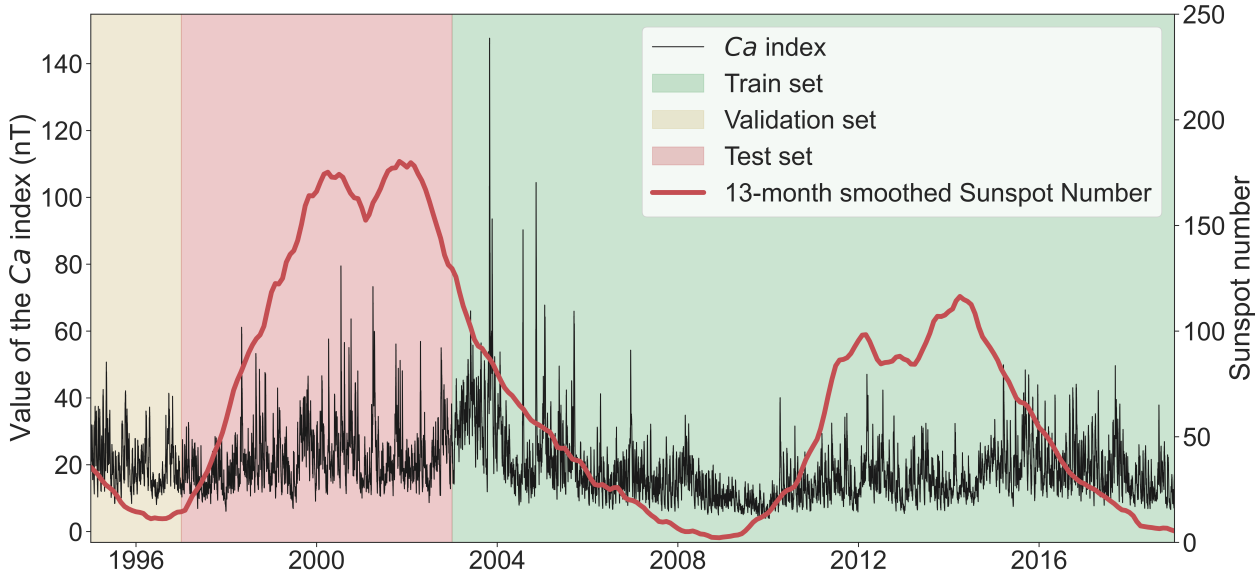

Fig. 2. Plot of the values taken by the *Ca* index between 1995 and 2018 included (black thin line). The training (green area), validation (yellow area), and test (red area) sets are highlighted. The 13-month smoothed Sunspot Number is also plotted as an indicator for the solar cycle (red thick line).

The train set is composed of 16 continuous years including the declining phase of one cycle and a full second cycle. The train set includes several extreme and even most extreme events, including the "Halloween storm" of November 2003 that reached a maximum value of *Ca* of 147.6 nT and was found to be the only 1-in-100 year event (in terms of *Ca* index) witnessed since the beginning

of the Space Era (Bernoux and Maget, 2020). The validation set is composed of a 2-year long period during a solar minimum. The test set is composed of 6 continuous years including the ascending phase, the maximum, and the beginning of the descending phase of a solar cycle. The test set includes intense and even extreme storms ($\geq$ 67 nT), which is a good step towards a fair evaluation of our model. The chosen split should ensure that our sets are representative enough of the space weather phenomena that can be observed through *Ca*.

To evaluate our model in an even more detailed way, we divide the test set into subparts corresponding to periods of disturbances induced on the one hand by ICMEs and on the other hand by Stream Interaction Regions (SIRs), including CIRs. For this purpose, we use the ICME database provided by Chi et al. (2016) and the SIR database provided by Chi et al. (2018). These databases include the time of beginning and time of ending for several ICME- and SIR-induced geomagnetic disturbances between 1995 and 2015 (2016 for SIRs). According to these databases, 212 SIRs and 204 ICMEs were observed in the near-Earth environment between 1997 and 2002 included. In our study, we define an ICME- (respectively SIR-) induced disturbance period as the time period during which an ICME- (respectively SIR-) induced geomagnetic disturbance has an influence on the dynamics of the *Ca* index. The beginning of the disturbance period is given by the beginning of the storm as indicated in the database. The ending of the disturbance period is given by adding $\tau = 4$ days to the ending of the storm as indicated in the database. We can hence evaluate our models using only the ICME- or SIR-induced disturbance periods and be able to better understand the accuracy of our forecasts. Table 1 summarises the number of data samples in each set and details the number of samples belonging to the disturbance periods.

Table 1. Number of data samples in each set, including the number of samples belonging to a disturbance period.

Data set	Total number of samples	Number of samples in a disturbance period		
		SIR-induced	ICME-induced	SIR- and ICME-induced
Training	139,512	> 77,710	> 28,047	> 4219
Validation	16,801	10,794	2,251	888
Test	51,841	27,776	24,058	5,407
Full	**208,154**	> **116,280**	> **54,356**	> **10,514**

Notes. The lists of SIR and ICME events we used end respectively in 2015 and 2016. Therefore, the number of samples in each disturbance period for the training set and the full set are actually greater than the ones reported in this table. This has no consequence in this study since we only split the test set according to the nature of the disturbance in order to evaluate the models.

2.3.2. Preprocessing the data

Before being fed into the neural network based model, the data are processed as follows:

- We interpolate the values of the *Ca* index in order to have hourly values instead of a value every 3 hours (this is meaningful since *Ca* is a very smooth time-integrated index and thus doing this interpolation changes neither the physics nor the statistics of the problem).
- Missing values in the other data sets are filled using SSA.

– Inputs are rescaled so that their mean is 0 and their standard deviation is 1. Outputs are rescaled to fit in the [0, 1] interval. The weights for performing the transformations are calculated only from the training data set in order not to include bias for validation and testing. This procedure is standard when working with recurrent networks.

3. Models and evaluation methods

In this section, we present the models used to predict the *Ca* index as well as the machine learning algorithms used in these models. We also describe the methods and measures used to evaluate the model.

3.1. Model description

The model developed in this study receives as input the past values of four solar wind parameters listed in subsection 2.1, namely the plasma bulk velocity (V_{sw}), the ion density (ρ), the southward component of the interplanetary magnetic field (IMF) B_z and the plasma temperature (T). Unlike other studies, we choose not to include the past values of the geomagnetic index as an input to the models because we position ourselves in an operational-like context. Indeed, even though the ISGI provides quick-look *aa* index values, reliance on two different data sources always presents a higher risk of data unavailability from one source, which is prejudicial when establishing a near-real-time forecasting service. Ideally for such a service, one would have both models (with and without historical geomagnetic indices as inputs), but this is out of the scope of this study and for clarity we only study one model in this paper. Here we use the 30 last days for each input (*i. e.* the 720 last hourly values). The inputs/outputs link can be summarised as follows:

$$\begin{pmatrix} V_{sw}(t-719) & \dots & V_{sw}(t-1) & V_{sw}(t) \\ \rho(t-719) & \dots & \rho(t-1) & \rho(t) \\ B_z(t-719) & \dots & B_z(t-1) & B_z(t) \\ T(t-719) & \dots & T(t-1) & T(t) \end{pmatrix} \longrightarrow \begin{pmatrix} Ca(t+1) \\ Ca(t+2) \\ \dots \\ Ca(t+n) \end{pmatrix}, \text{ where } n \text{ is the forecast horizon.}$$

In section 4 we will analyse the results for a model trained and tested with a forecast horizon $n = 24$ hours.

Our main model is a neural network-based model. It consists of a single layer Long-Short Term Memory network (LSTM) combined with a linear fully-connected feed-forward (FCFF-NN) layer. LSTMs are a type of recurrent neural networks first introduced in Hochreiter and Schmidhuber (1997). LSTMs were created to address problems involving sequentially-structured data such as time series or natural language. In particular, LSTMs possess two internal memory states that are designed to help address the gradient vanishing issue that occurs when handling long sequences (Hochreiter, 1998). For an in-depth understanding of deep learning methods, including recurrent and LSTM networks, the reader is referred to the above-mentioned papers as well as to reference textbooks such as Goodfellow et al. (2016).

Our model is summarised in Figure 3.

Let us summarise the functioning of the LSTM network here. For each sample corresponding to a time step $t - p$, the LSTM cell is fed with our solar wind parameters $\mathbf{x_{t-p}}$ and the two memory

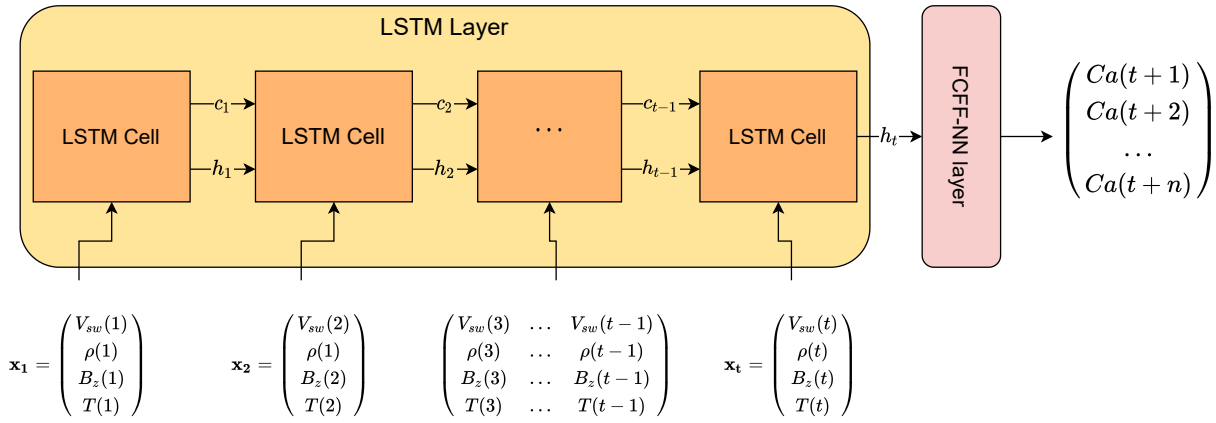

Fig. 3. Simple scheme representing the LSTM-based model to forecast the values of the *Ca* index up to *n* hours in advance. The mechanism inside the LSTM cell was voluntarily not detailed.

states computed at the previous time step: the hidden state h_{t-p} and the cell state c_{t-p}. The LSTM cell processes and transforms the input and updates its hidden state and cell state (now h_{t-p+1} and c_{t-p+1}) using three "gates": the input gate, the output gate, and the forget gate. To put it in simple words, the LSTM cell decides which information from the past is "worth" being kept, forgotten, or updated according to the last input. The latest memory states are again fed to the LSTM cell along with the solar wind parameters at the next time step $\mathbf{x_{t-p+1}}$. After all time steps have been given to the network, the LSTM layer outputs the final hidden state h_t that serves as the input to the FCFF-NN layer, which itself outputs the $t + 1$ to $t + n$ next values of *Ca*, n being the forecast horizon.

Let us note that LSTMs have already demonstrated a good efficiency on geomagnetic index prediction problems (see e.g. Gruet et al., 2018; Chakraborty and Morley, 2020; Laperre et al., 2020).

Since this is the first study that focuses on the forecast of the *Ca* index, there is no immediate baseline for us to compare our model to. The usual baseline used in such a situation is the "persistence model" (also known as the "naive model"), which simply consists in assuming that the predicted value is the same as the last observed value. However, that baseline cannot be pertinently used here as we do not include the past values of *Ca* index among the inputs to our model. That is why we have also trained a simple linear regression model to forecast the *Ca* index from the same solar wind parameters as with the neural network-based model, with the notable exception that the baseline linear model only uses the last value for each solar wind parameter as input (and not several past values as with the neural network-based model).

3.2. Training and parameters of the model

Our model was trained using the classical backpropagation method (Rumelhart et al., 1986). The optimisation method used is the Adam algorithm (Kingma and Ba, 2017). We have used a learning rate $lr = 1 \times 10^{-4}$ that is halved a first time after epoch 15 and a second time after epoch 50. The loss function is the mean-square error (MSE). The parameters of the model were hand-picked

using cross-validation and iteration. We list below the main parameters of our model and some implementation choices so that the replicability of our results is made easier. Let the reader be advised that even after changing some of these parameters (e.g. in order to reduce the computational cost) it is possible to obtain very similar results.

- The LSTM cell state has dimension 256.
- The LSTM layer is mono-directional.
- We use L2-regularisation with weight 1×10^{-5}. L2-regularisation consists in adding the squared sum of the network's weights (with a multiplicative constant) to the loss function in order to avoid overfitting.
- Size of each mini-batch: 256.
- The training is done with 120 epochs and with early stopping. Early stopping consists in stopping the training of the network as soon as clear signs of overfitting are observed.

The model was developed using the PyTorch (v1.9) library for Python (Paszke et al., 2019).

3.3. Detection of events

Our models as described above offer predictions in the form of a regression problem. However, it is often more useful for an end-user in a decision-making context to benefit from a predictive alert system. Such a (binary) predictive alert system can be built from our (regression) models with the following method: if we predict that *Ca* will exceed a given threshold value during the next t hours then we issue an alert (class 1), if we predict that we stay below this threshold then we issue no alert (class 0). The only difficulty lies in the choice of a suitable threshold.

In our example, we will choose a threshold value based as much as possible on operational criteria. The threshold must be meaningful to the end-user, *i.e.* the triggering of an alert must correspond to a situation for which the operator is expected to make a decision or take an action. As the *Ca* index represents the filling state of radiation belts with high-energy electrons, we will choose a *Ca* threshold associated with a non-negligible risk of damage due to surface charging.

Figure 4 in Bernoux and Maget (2020) shows that *Ca* index has a quite high correlation coefficient ($R \approx 0.83$) with the dynamics of the integrated $E \geq 30$ keV electron flux at $L* \approx 6$. Moreover, Matéo-Vélez et al. (2018) shows that the risk of damage due to surface charging for a spacecraft in geostationary orbit (i.e. at $L* \approx 6$) is well correlated with the $10 \leq E \leq 50$ keV electron flux when the latter is greater than 1×10^8 cm^{-2}s^{-1}sr^{-1}. A day during which the $10 \leq E \leq 50$ keV electron flux always stayed above this value has a minimum daily fluence of 8.64×10^{12} cm^{-2}sr^{-1}. From this value we define a fluence threshold equals 8×10^{12} cm^{-2}sr^{-1}.

We then tried and find a *Ca* threshold that gives the highest correlation between the monthly exceedances of the electron fluence and the monthly exceedances of the *Ca* threshold (using the daily *Ca* maximum). For the daily fluences, we have taken data provided by the Magnetospheric Plasma Analyzer (MPA) instrument onboard the Geosynchronous Equatorial Orbit (GEO) LANL 1991-80 spacecraft between 1997 and 2006 for the energy range 35-46 keV (McComas et al., 1993).

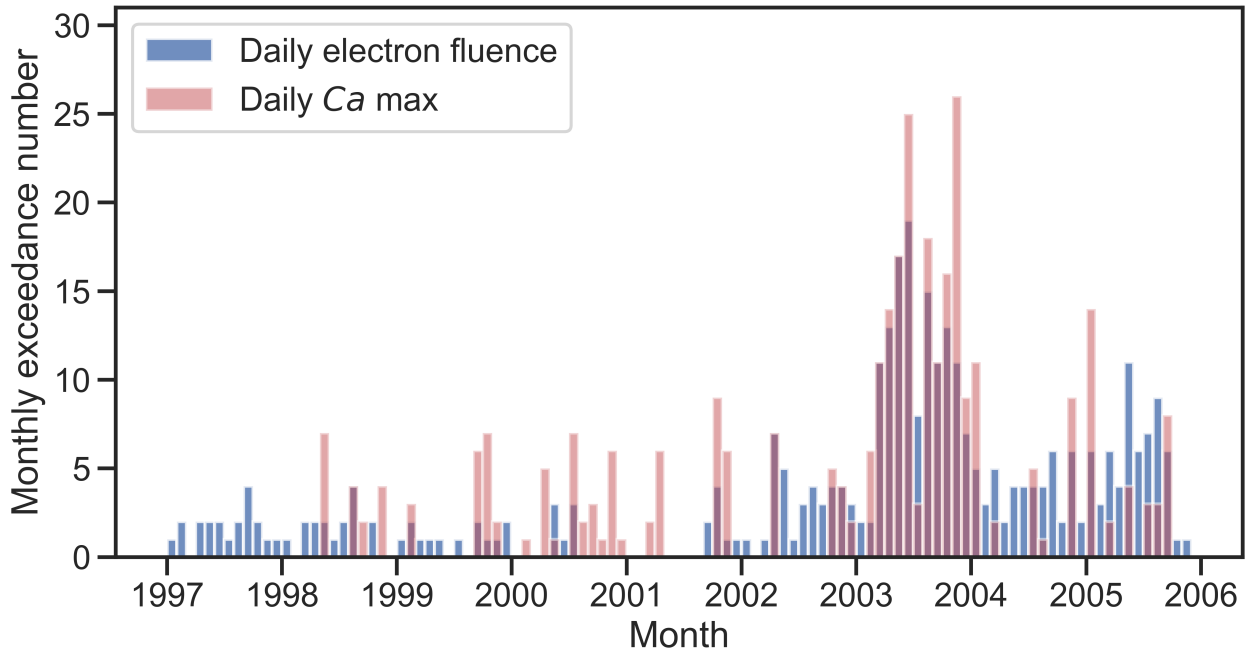

Fig. 4. Count of days per month for which LANL 1991-80/MPA instrument measured a daily $10 \leq E \leq 50$ keV electron fluence above 8×10^{12} cm^{-2}sr^{-1} along with the count of days per month for which the daily *Ca* max was above 38 nT.

It was found that the number of monthly fluence exceedances is best correlated with the monthly *Ca* exceedances when the *Ca* threshold is $Ca_{\text{threshold}} = 38$ nT. This is also illustrated in Figure 4.

It should be noted in hindsight that the *Ca* value of 38 nT corresponds approximately to the 0.95 percentile of all *Ca* values, which seems statistically satisfactory. Indeed, it is a value that is therefore rare enough to make a credible and useful alert threshold (an operator would probably not want to receive an alert when the *Ca* value only exceeds the median, for example). But it is also a value that is not too high, which allows better learning for the neural network (indeed, the higher the threshold, the fewer samples we have to train and evaluate the model). Let us also insist on the fact that this threshold value used to define our binary classes in our study is only an example, and that depending on the effect considered (internal charging, surface charging, singular events, etc), the orbit considered, or even the satellite considered (and thus its structure) it would be more interesting to use other thresholds, and probably to increase the number of classes.

3.4. Model evaluation

In this subsection we describe the measures used to evaluate the forecast performance of our models.

3.4.1. Regression metrics

Since our problem is designed as a regression problem we first evaluate our model using two very common regression metrics: the root-mean-square error (RMSE) and the Pearson (linear) correla-

tion coefficient (R). Let us define y_i the real observed values and $\bar{y}_i$ the values forecast by a model for $i \in 1, ..., N$, N being the number of samples.

- The RMSE is a measure of the global accuracy of the model, with more emphasis put on higher values (e.g. here the emphasis is on periods of more intense geomagnetic activity). A lower RMSE means a more accurate forecast. The RMSE is given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i - y_i)^2} \tag{2}$$

- The Pearson correlation indicates if the forecast values globally follow the same trends as the real values. The Pearson correlation ranges between 0 and 1 (higher is better). It is given by:

$$\mathrm{R} = \frac{\mathrm{Cov}(\bar{y}_i, y_i)}{\sqrt{\mathrm{Var}(\bar{y}_i) \times \mathrm{Var}(y_i)}} \tag{3}$$

We also use a normalised version of the RMSE (NRMSE), which allows for better comparison of data sets with different scales. The NRMSE is obtained by dividing the RMSE by the mean value of the observed y_i. It is given by:

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i - y_i)^2}}{\frac{1}{N}\sum_{i=1}^{N} y_i} \tag{4}$$

Both RMSE and Pearson correlation are widely used in the geomagnetic indices forecasting literature (e.g. in Lazzús et al. (2017); Tan et al. (2018); Gruet et al. (2018); Sexton et al. (2019)). However, these metrics do not capture the full performance of a model in all situations. Indeed, these metrics indicate overall trends. Most of the time geomagnetic activity is fairly quiet, so quiet periods will weigh much more heavily on the evaluation metrics than periods of high activity, thus creating a bias. While it is very interesting for a satellite operator to be able to accurately predict quiet periods, it is also very important to be able to accurately predict periods of geomagnetic disturbance. This type of bias can be partially counterbalanced by taking adapted test sets, as we have done in Section 2.3. In the following subsections, we describe two other methods for evaluating the predictions that allow us to better capture other types of behaviours.

3.4.2. Measuring time lags

Some studies, such as Wintoft and Wik (2018) and Laperre et al. (2020), highlight the fact that some forecasting models, that display a great RMSE or Pearson correlation, actually fail to reliably forecast high disturbance periods in advance. Laperre et al. (2020) shows that some prediction models exhibit systematic time lags between the observed time series and the predicted time series. This systematic time lag would most often be of the order of magnitude of the model's prediction horizon. This would indicate that the model in reality would fail to predict a disturbance before it has actually been observed, which is of very limited interest to an operator.

To quantify this behaviour Laperre et al. (2020) use the Dynamic Time Warping (DTW) algorithm, which measures the time difference between two time series (Berndt and Clifford, 1994). By applying this algorithm to the observed series and the predicted series shifted successively by several consecutive time steps the authors are able to determine the extent of the systematic lag. Nonetheless in our study we do not use the exact same approach but a very similar one. Indeed, the main drawback of the DTW method is that for a given prediction horizon n, it requires circa n^2 iterations of the DTW algorithm with different time shifts to accurately assess the systematic time lag. Besides, the computational complexity of the DTW algorithm is high even with now modern methods to fasten the computation of the DTW measure (e.g. Gold and Sharir, 2018). This is why we use instead the Temporal Distortion Mix (TDM).

The Temporal Distortion Mix is a metric proposed in Vallance et al. (2017) to characterise the propensity of a time series to be late or early relative to a reference series. This metric is also based on the DTW algorithm. Based on this algorithm, Frías-Paredes et al. (2016) proposes the Temporal Distortion Index (TDI), which indicates to what extent the two time series are systematically (or not) late (or early). Unlike the approach proposed by Laperre et al. (2020), the TDI does not indicate the value of a possible systematic time lag, but whether the two time series exhibit this type of behaviour and to which extent. In return, there is no need for several computations of the DTW measure as only one (per forecast horizon) is sufficient to get the TDI. Guen and Thome (2019) have even suggested that the TDI could be used as a part of the loss function when training a neural network but this is out of the scope of our paper.

To obtain the TDM, the TDI is decomposed into two components, which characterise the lateness and the advance, so that TDI = TDI_{adv} + TDI_{late}. The TDM is then given by:

$$\text{TDM} = 1 - 2 \times \frac{\text{TDI}_{adv}}{\text{TDI}} \tag{5}$$

The TDM is hence a normalised version of the TDI. It ranges between -1 and 1. Let $\mathbf{s_1}$ and $\mathbf{s_2}$ be two time series.

- if $\text{TDM}(\mathbf{s_1}, \mathbf{s_2}) = -1$ then $\mathbf{s_1}$ is systematically in advance compared to $\mathbf{s_2}$
- if $\text{TDM}(\mathbf{s_1}, \mathbf{s_2}) = 1$ then $\mathbf{s_1}$ is systematically late compared to $\mathbf{s_2}$
- if $\text{TDM}(\mathbf{s_1}, \mathbf{s_2}) = 0$ then both time series are temporally aligned

For instance, the TDM between a given time series and its corresponding naive forecast is always 1. A good forecast is hence a forecast that has a TDM close to 0. The TDM is a very interesting evaluation measure since it only requires one run of the DTW algorithm and it is possible to compare the TDM between several forecasts (e.g. several forecast horizons). The TDM was first introduced in a study dealing with the topic of solar irradiance forecasting, which is also a time series forecasting problem that shares structural similarities with ours.

3.4.3. Evaluation of the classification-based alert system

As we have already established, in an operational context in space weather it is important not only to have regression type predictions but also to have warning systems based on class predictions. In Section 3.3 we discussed how to transform our regression problem into a binary classification

problem (with a threshold of $Ca_{threshold}$ = 38 nT). In order to evaluate this derived alert system, we use several metrics and measures. TP, FP, FN and TN are the true positive, false positive, false negative, and true negative counts.

- the precision: it is the ratio of issued alerts that match a true threshold excess. It gives an indication of how relevant the issued alerts are. It ranges between 0 and 1. Higher is better. It is given by:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

- the recall: it is the ratio of true threshold exceedances that match an issued alert. It gives an indication of the ability to issue relevant alerts. It ranges between 0 and 1. Higher is better. It is given by:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

- the F_{score}: it is the harmonic mean of precision and recall. It ranges between 0 and 1. Higher is better. It is given by:

$$F_{score} = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

- the False Alarm Rate (FAR): it is the ratio of nonevents for which an alert was issued. It gives an indication of the tendency to issue irrelevant alerts. It ranges between 0 and 1. Lower is better. It is given by:

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{9}$$

- the threat score (TS): it gives an indication of how well true threshold exceedances were forecast, penalising both false alarms and false negatives. It ranges between 0 and 1. Higher is better. It is given by:

$$\text{TS} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{10}$$

- the Heidke skill score (HSS): it could be seen as a generalised skill score, giving the overall accuracy of the model against that of a random model. It ranges between -1 and 1. Higher is better, 0 denotes no skill. It is given by:

$$\text{HSS} = \frac{2 \times (\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{(\text{TP} + \text{FN})(\text{FN} + \text{TN}) + (\text{TP} + \text{FP})(\text{FP} + \text{TN})} \tag{11}$$

- The percentage of threshold-exceedance periods for which the model actually issues an alert before the threshold was exceeded (*i.e.* the number of active periods that were forecast before they started and not only forecast after the threshold was exceeded for the first time). This is not a classical metric, but perhaps one of the most useful ones here, since this gives an indication of how well the model is able to forecast disturbance periods before they happened, not including the performance of the model once the disturbance period has already started. Let us note that there are 42 disturbance onsets (above the threshold Ca = 38 nT) in the test set.

4. Results and discussion

4.1. Regression results

The regression results obtained with the baseline model and the LSTM-NN model are presented in Table 2. Firstly, we can see that the classical metrics (RMSE, R) give much better values with the LSTM-NN model than with the linear baseline. For a time horizon of 3 hours, the RMSE with the LSTM-NN is about 3.1 times lower than with the baseline (2.62 instead of 8.13), and for a time horizon of 24 hours this ratio is 2.0 (4.17 instead of 8.16). This is an additional indication to the fact that LSTM-NN networks are efficient for understanding the solar wind-magnetosphere coupling. The RMSE values should be put into perspective with the statistical distribution of the *Ca* index, which over the test period has a variance of 8.9 nT and an interquartile range of 10.5 nT. This comparison allows us to state that the RMSE values are satisfactory, especially for a model that does not include the *Ca* index among its inputs. We also find that the Pearson correlation values are quite high (≥ 0.9 for all test sets up to a time horizon of 18 hours, instead of ≤ 0.65 with the baseline), which is very satisfactory.

The TDM gives values ≤ 0.2 for a time horizon of 3 hours and up to 6 hours, for test sets based on periods of disturbance. This indicates that up to about 6 hours, our forecasts are well aligned in time with the target values. Beyond that, the TDM value increases up to 0.60 for a 24 hour time horizon with the full test set, indicating that there is an almost systematic delay between the predicted values and the target values.

Unsurprisingly, the values of the conventional metrics all degrade as the time horizon increases. This degradation (increase for RMSE and TDM, decrease for the Pearson correlation) appears to be slow and smooth, as shown in Figure 5. However, for this reason, it becomes difficult to tell from these metrics alone from which time horizon the model is no longer operationally valid.

We also observe that, in general, the LSTM-NN model performs better during periods of SIR-induced disturbances than during periods of ICME-induced disturbances. For a time horizon of 3 hours, the RMSE is 1.4 times higher for the ICME-induced period than for the SIR-induced period, which is far from negligible. Figure 6 shows several examples of forecasts for two geomagnetic storms: one induced by an ICME and the other by a SIR, the same storms already shown in Figure 1. This figure shows the forecast values for 4 different time horizons (3, 6, 12, and 24 hours) made with both the LSTM-NN model and the linear baseline model. We also indicate the TDM values calculated corresponding to each forecast (the values for the baseline are in brackets). It is clear from this figure that the neural network-based model outperforms the linear model, as already indicated by the evaluation measures for the regression problem. In these examples, the dynamics of the storm appear to be well captured, and the forecast values are indeed close to the observed values, as indicated by the RMSE. Furthermore, it becomes apparent that the negative TDM values measured with the linear model are due to the fact that the model has difficulty correctly modeling the decay phase of a storm, which decreases too fast and hence appears "ahead" in comparison to the true series.

Besides, the fact that the predicted (with the LSTM-NN model) and observed time series show a time delay as the time horizon increases is evident in these examples. It would appear that this time shift is more pronounced during the beginning of the disturbance period than during the decay phase of the storm, which in the SIR-induced storm example remains well predicted even 24 hours

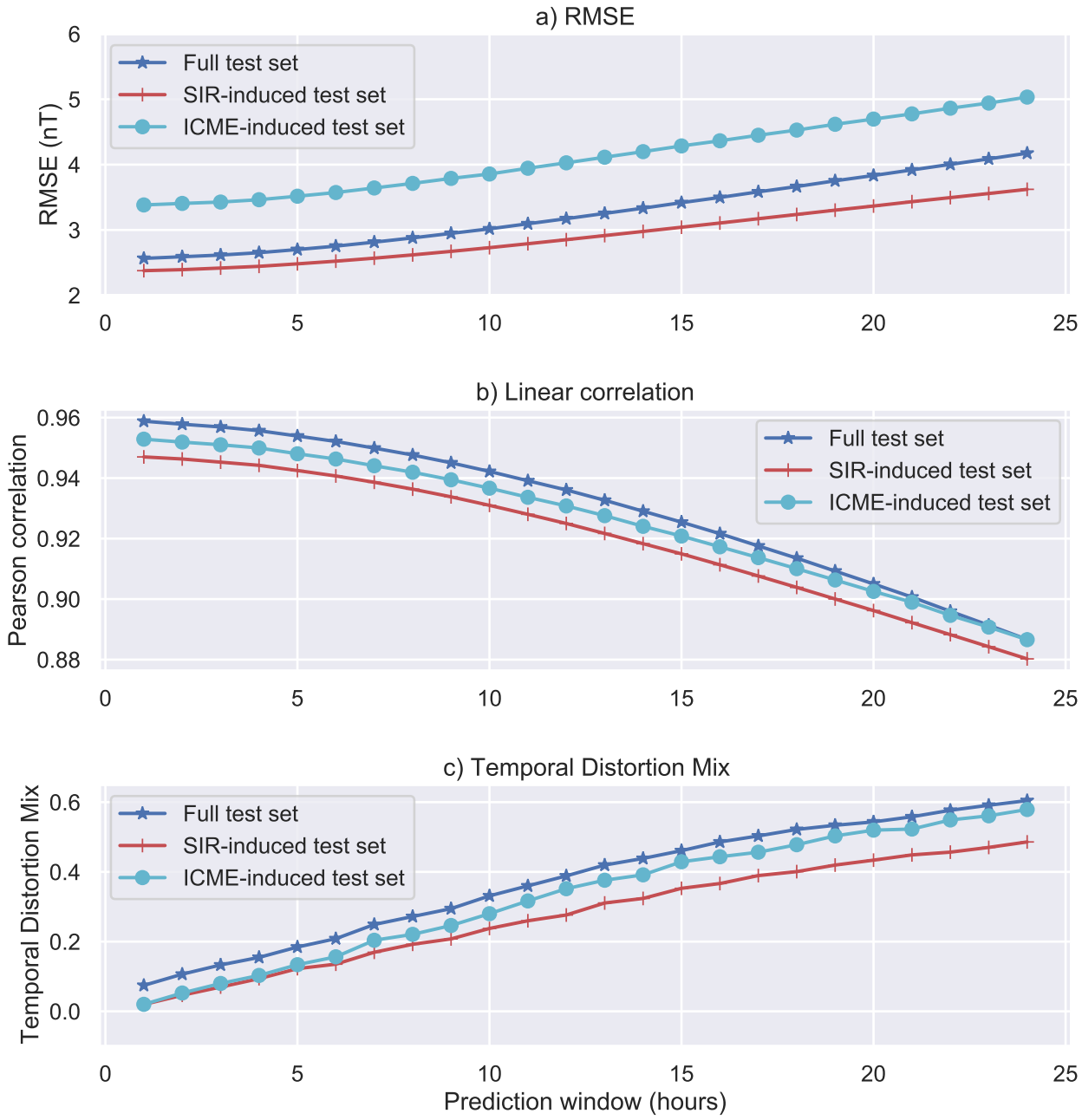

Fig. 5. Evaluation of the LSTM-NN model with three measures (RMSE, R and TDM) for values of time horizon ranging from 1 hour to 24 hours. Three evaluation sets (full test set, SIR-induced set and ICME-induced set) were used.

in advance. We should be able to better quantify this behaviour using the measures for the evaluation of the classification problem.

The difference of performance between ICME-induced and SIR-induced storms could hence at least partly be explained by the fact that *Ca* increases more rapidly during ICME-induced disturbances. As indicated by the TDM values (and as we will see below with the classification measures),

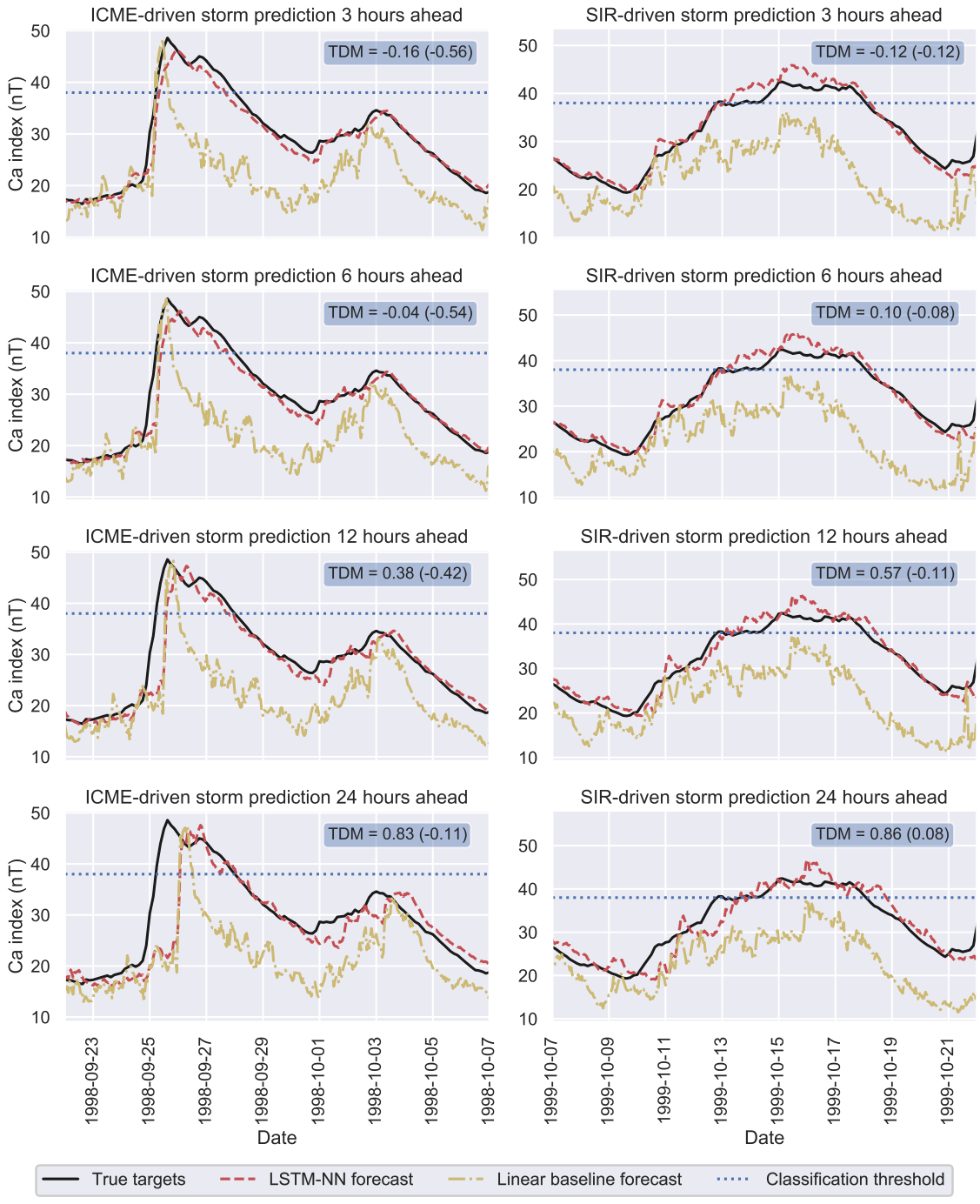

Fig. 6. Example of forecasts obtained with the LSTM-NN model and the linear model during two geomagnetic storms, the first one (left-hand side) being an ICME-driven storm and the second one being a SIR-driven storm (right-hand side). 4 different forecast horizons were used (3, 6, 12 and 24 hours). The value of *Ca* used for the binary classification is the blue dotted line, given as a landmark. For each prediction, the corresponding TDM value is given (the TDM values for the baseline forecasts are given in brackets).

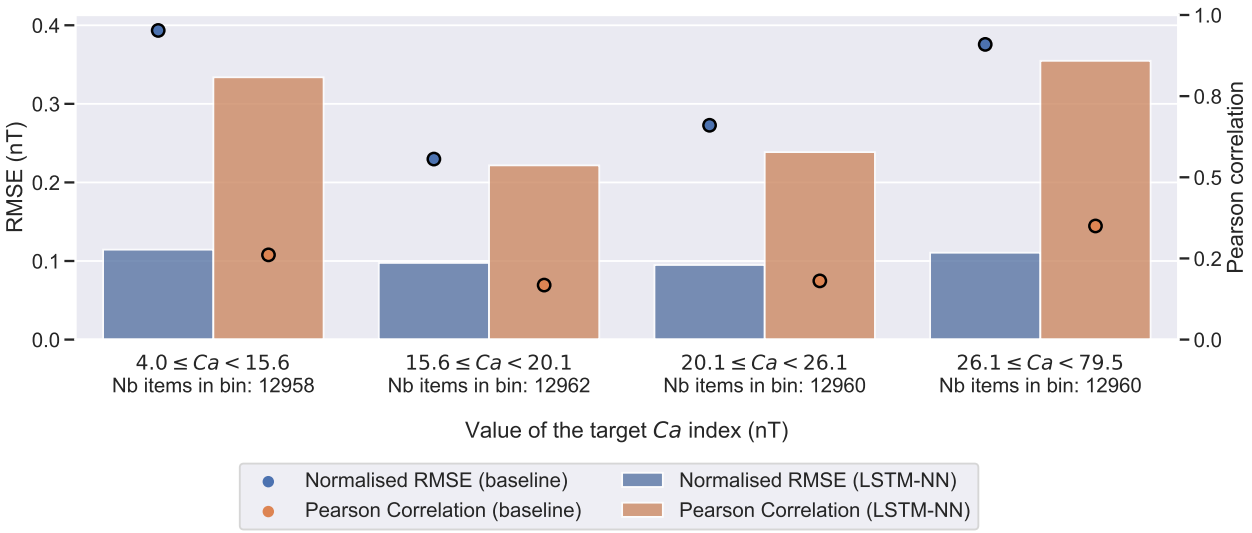

Fig. 7. Normalised RMSE and Pearson correlation of the 3-hour ahead predicted values versus binned observed values. Each bin contains a quarter of the total observations in the test set.

the LSTM-NN model seems to be under-performing during the initial phase of a disturbance. Since during SIR-induced disturbances the initial increase is slower than during ICME-induced disturbances, the RMSE during the beginning of the disturbance period should be lower in the first case, which contributes to the overall RMSE being lower for the SIR-induced test set than for the ICME-induced test set.

Figure 7 shows the Normalised RMSE (NRMSE) and the Pearson correlation for forecasts with a time horizon of 3 hours, after binning the target values into quarters containing more-or-less the same number of items. Here we use the NRMSE since we are comparing the forecasts for different scales of *Ca* values, thus using the RMSE for the comparison would be like comparing apples and oranges. It appears that the LSTM-NN model gives stable NRMSE values when *Ca* increases, which shows that the model is performing similarly not only when *Ca* is low, but also when it reaches higher values, unlike the baseline. The Pearson Correlation for both models decreases when *Ca* is between the first and the third quartile. This is most probably due to the choice of bins matching the quartiles of *Ca*. Indeed, the range of *Ca* values in these bins is less than 6 nT, i.e. of the order of only twice the RMSE. Therefore, it is not surprising that the spread of predicted values over such a small range of observed values makes the linear correlation in these bins weaker. To summarise, this figure shows us that the model gives stable results and is still a much better model than the linear model for the whole distribution of *Ca* values, with a notable improvement for high values of *Ca*.

4.2. Classification results

The classification results are given in Table 3 and Figure 8. For a time horizon of 3 hours, nearly 85% of the alerts issued were true positives, while 84% of the threshold exceedances were detected. For a time horizon of 24 hours, these numbers rise and fall respectively to 87% and 73%. The fact that the precision increases with the time horizon is due to the definition of our binary classes.

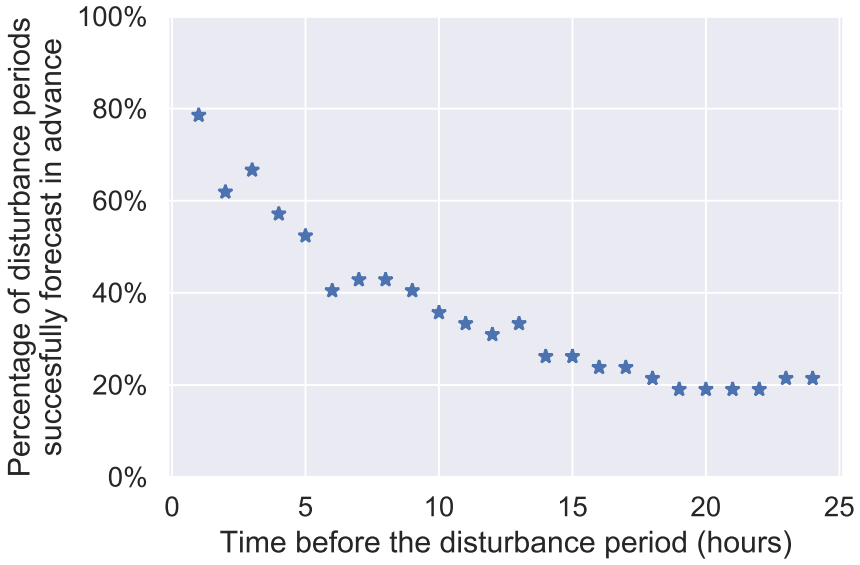

Fig. 8. Percentage of times the 24h-binary classification problem was correctly forecast during quiet periods previous to a threshold exceedance depending on how much time (from 1 hour to 24 hours) there was left before the exceedance.

Indeed, we are trying to forecast if the threshold will be exceeded at any given time in the next t hours (and not at a precise given time). In our case, as the threshold increases, the model forecasts less often true and false positives and more false negatives. That is why the precision increases somewhat counter-intuitively. This highlights the need for several evaluation methods in order to obtain a more exhaustive idea of the true performance of the model. Let us note that the F_{score}, which is the harmonic mean of precision and recall, decreases from 0.84 (for a time horizon of 3 hours) to 0.80 (for a time horizon of 24 hours), further indicating that the model performs better for shorter time horizons.

It is difficult to argue at what percentage of precision and recall the model becomes satisfactory. In absolute terms, correctly predicting more than two out of three periods of disturbance while making only $\approx 25\%$ false positives might seem to be a satisfactory target. However, depending on the economic constraints due to spacecraft operation this could be largely insufficient. Here we cannot definitively conclude about the absolute quality of our model but only about criteria that would be defined by an operator and that depend on each space mission or on the targeted objective. It should be noted, however, that the score values are also quite high, especially for the HSS. In absolute terms, these values are rather difficult to interpret and should serve above all as a point of comparison for possible future studies focusing on the forecast of similar physical quantities.

A result that is easier to interpret and that gives user-friendly information is the percentage of disturbance periods forecast in advance, given in Figure 8. To obtain this figure we calculated the percentage of times and how long before the model was able to correctly answer the question: "will the threshold be exceeded during the next 24 hours?" Therefore here we are only interested in the model's ability to predict the beginning of a period of disturbance (without taking into account the continuation of such a period). It appears that the model is able to answer this question correctly slightly less than 80% of the time 1 hour before the threshold is exceeded. This percentage remains

above 50% up to 5 hours before the threshold is exceeded. Between 6 and 9 hours prior to a threshold exceedance, around 40% of the disturbance periods were correctly forecasted. Less than 25% of the threshold exceedances were forecast at least 15 hours in advance. This shows that even though 73% of the total exceedances were detected somewhere between 1 hour and 24 hours before they happened, only less than one out of two disturbance periods were detected 6 hours before they happened and less than one out of four were detected 24 hours before they happened. This is a much more significant measure of the operational nature of our model and confirms the point we made earlier about the difficulty of predicting the onset of a geomagnetic storm.

4.3. Discussion

In fact, the above-mentioned results are not very surprising since our models rely on solar wind parameters measured close to the Earth. Consequently, the temporal hindsight to predict the dynamics of radiation belts is small. This is reflected in the TDM measurements which indicate that the forecasts are globally very well temporally aligned with the observations for forecast horizon values shorter than 6 hours, which corresponds approximately to the reaction time of the geomagnetosphere interacting with a disturbance arriving near Earth. We can therefore deduce on the one hand that our model seems to be in agreement with the physics of the problem. But on the other hand, if we do not change the nature of our inputs, the same physics stops us from having good operational performances for greater prediction horizons.

Moreover, it seems delicate to find a limit to the prediction horizon for our model, beyond which it is possible to state definitively that the model is no longer operational. As mentioned above, this depends on the needs of an operator. In the absence of threshold values that could serve as landmarks for metrics such as precision or recall, we can only guess. One way to do this would be to consider the percentage of storms predicted in advance. If we take a threshold of 50%, then the operational prediction horizon limit of our model is 5 hours. With a threshold of 75% then our operational prediction horizon limit is only 1 hour. Another method would be to take into account the TDM. With an arbitrary threshold of 0.2, the prediction horizon limit of our model is 6 hours, whereas with a threshold of 0.1 the horizon limit is only 1 hour for the full test set, but 4 hours during SIR-induced disturbance periods and 3 hours during ICME-induced disturbance periods.

A limit of 6 hours was found in other papers dealing with the forecasting of the *Dst* index (Lazzús et al., 2017; Gruet et al., 2018). Some studies that aim at forecasting the *Kp* index, such as Tan et al. (2018); Sexton et al. (2019), claim to be able to forecast the *Kp* index up to 24 hours in advance. It would be interesting to assess the operational performance of the models presented in these papers with the TDM and by evaluating only the ability to predict the onset of a storm, in order to have a more comprehensive understanding of their actual effectiveness in operational contexts. Let us insist, however, on the fact that the difficulty for long prediction horizons lies at the beginning of the storm and not in its continuity because the accumulation of energy makes it possible to find a link between the solar wind parameters and the geomagnetic indices even after 6 hours of course. This is particularly the case with a time-integrated index such as *Ca*, which allows for good overall forecast performances up to 24 hours in advance.

It might be tempting to compare our results to the results presented in e.g. Forsyth et al. (2020) where the authors present a model to forecast the GOES-15 $\geq$ 2MeV electron fluxes from solar wind data and also evaluate their model with classification measures. For instance, one of their models

(when maximising the average Receiver Operating Characteristic score) for a time horizon of 6 hours gives a hit rate (or precision) of 0.75 whereas for the same time horizon ours give a higher hit rate of 0.87. However, this comparison does not stand because we are not focusing on the same energy range and our model does not use the same classification thresholds and criteria. Indeed, here we answer the question: will the threshold be exceeded somewhere in the next t hours? In Forsyth et al. (2020) the question is: will the threshold be exceeded in exactly t hours? We have chosen to approach the problem in this way because we believe that a warning system defined in this way is more useful, especially if we ask this question for several time horizons t. Yet this is an arbitrary choice and it could be argued otherwise. We wanted to stress here that, as highlighted in Camporeale (2019), comparing the performance of one model relative to another is not straightforward, and one should be cautious when doing it.

Table 2. Evaluation of the NN-based and the baseline models in the context of the regression problem. The model was evaluated with the full test set and also with the SIR-induced test set and the ICME-induced test set.

Time horizon	RMSE (nT)			R			TDM		
(hours)	Full	SIR	ICME	Full	SIR	ICME	Full	SIR	ICME
3	**2.62** (8.13)	**2.42** (6.37)	**3.43** (11.18)	**0.96** (0.63)	**0.95** (0.64)	**0.95** (0.64)	**0.13** (-0.42)	**0.07** (-0.30)	**0.08** (-0.55)
6	**2.75** (8.08)	**2.52** (6.37)	**3.57** (11.10)	**0.95** (0.64)	**0.94** (0.64)	**0.95** (0.65)	**0.21** (-0.37)	**0.14** (-0.25)	**0.16** (-0.51)
9	**2.95** (8.05)	**2.67** (6.39)	**3.79** (11.05)	**0.95** (0.64)	**0.93** (0.64)	**0.94** (0.65)	**0.39** (-0.32)	**0.21** (-0.21)	**0.25** (-0.49)
12	**3.17** (8.05)	**2.85** (6.43)	**4.03** (11.01)	**0.94** (0.64)	**0.92** (0.64)	**0.93** (0.65)	**0.39** (-0.28)	**0.28** (-0.16)	**0.35** (-0.46)
15	**3.42** (8.05)	**3.04** (6.48)	**4.29** (10.97)	**0.93** (0.64)	**0.91** (0.63)	**0.92** (0.64)	**0.46** (-0.23)	**0.35** (-0.12)	**0.43** (-0.41)
18	**3.66** (8.08)	**3.24** (6.54)	**4.53** (10.94)	**0.91** (0.63)	**0.90** (0.63)	**0.91** (0.64)	**0.52** (-0.17)	**0.40** (-0.06)	**0.48** (-0.35)
21	**3.92** (8.11)	**3.43** (6.59)	**4.78** (10.92)	**0.90** (0.62)	**0.89** (0.62)	**0.90** (0.63)	**0.56** (-0.13)	**0.45** (-0.02)	**0.52** (-0.30)
24	**4.17** (8.16)	**3.62** (6.65)	**5.03** (10.90)	**0.89** (0.62)	**0.88** (0.61)	**0.89** (0.63)	**0.60** (-0.10)	**0.49** (0.03)	**0.58** (-0.29)

Notes. The results obtained with the NN-based model are given in bold. The results obtained with the baseline are given in brackets.

Table 3. Evaluation of the NN-based and the baseline models in the context of the classification problem.

Time horizon (hours)	Precision	Recall	F_{score}	FAR	Threat score	Heidke Skill Score
3	**0.85** (0.64)	**0.84** (0.07)	**0.84** (0.12)	**0.010** (0.002)	**0.73** (0.06)	**0.83** (0.11)
6	**0.85** (0.66)	**0.83** (0.07)	**0.84** (0.12)	**0.010** (0.002)	**0.73** (0.07)	**0.83** (0.11)
9	**0.86** (0.67)	**0.82** (0.07)	**0.84** (0.12)	**0.009** (0.002)	**0.72** (0.07)	**0.83** (0.11)
12	**0.87** (0.68)	**0.80** (0.07)	**0.83** (0.13)	**0.009** (0.002)	**0.71** (0.07)	**0.82** (0.12)
15	**0.87** (0.69)	**0.79** (0.07)	**0.83** (0.13)	**0.009** (0.002)	**0.70** (0.07)	**0.81** (0.12)
18	**0.87** (0.69)	**0.77** (0.07)	**0.82** (0.13)	**0.009** (0.002)	**0.69** (0.07)	**0.80** (0.12)
21	**0.87** (0.70)	**0.75** (0.07)	**0.81** (0.13)	**0.009** (0.002)	**0.68** (0.07)	**0.79** (0.12)
24	**0.87** (0.69)	**0.73** (0.07)	**0.80** (0.13)	**0.009** (0.002)	**0.66** (0.07)	**0.78** (0.12)

Notes. The results obtained with the NN-based model are given in bold. The results obtained with the baseline are given in brackets.

5. Conclusion

In this study, we propose a recurrent network-based approach to forecast the fairly new geomagnetic index *Ca*. The main reason for focusing on this index is that this index is well correlated with the high-energy electron fluxes in the radiation belts and could hence be used as an indicator for their state of filling, without the drawbacks inherent to measuring *in-situ* fluxes with spacecrafts.

The implementation choices made in this paper were made by keeping in mind an operational context. These choices include the geomagnetic index to be forecast, the inputs used in our models, and the whole evaluation methodology. To this end, we have highlighted the importance of choosing statistically and physically representative train and test sets. We have also stressed the need to use adequate measures to evaluate the model, since classical metrics such as the RMSE or the Pearson correlation are not able to give an exhaustive report on the performance of the model, in particular during disturbance periods. That is why we use the Temporal Distortion Mix to measure the tendency for a forecast to be late or in advance in regards to the true observations.

We also transform the forecast problem from a regression problem to a binary classification one. The choice of the threshold used to define the binary classes was made taking into account the risk for GEO spacecrafts to suffer damage from the surface charging effect. The evaluation of the binary classification forecasts shows that even though the regression measures seemed great, the network does not show outstanding performance when it comes to forecasting the onset of a disturbance period. This is most certainly due to the spatial (and hence temporal) proximity between the solar wind parameters used as inputs and the geomagnetosphere. In order to improve the forecast results for time horizons of 12 hours, 24 hours, and beyond it could be interesting to go back to the Sun and use data originating from solar imaging as inputs to a model. This topic will be the main focus of future studies. For now, even though the measures are good and much better than the linear baseline, it would be difficult to claim that this model is fully adequate for use in an operational situation. This would require at least an assessment of the model's ability to predict extreme events, which will be the subject of future studies. However, with this study, we have already taken a first great step towards this goal.

Other possibilities that remained out of the scope of this study are the use of probabilistic forecasts (as done with other indices e.g. in Chandorkar et al., 2017; Chakraborty and Morley, 2020) or grey-box models. This paper being the first one dealing with the topic of forecasting the *Ca* index, we voluntarily kept those possibilities aside for the sake of clarity and so as not to dilute the purpose of this study. However, we acknowledge that these are important avenues to explore, which will be done in future studies.

Acknowledgements. The authors would like to thank the anonymous reviewers for their insightful suggestions and comments, which helped improve the overall quality of the paper. The authors are thankful to the NOAA-POES for online data access available on the CDAweb (at http://cdaweb.gsfc.nasa.gov/). The results presented in this paper rely on geomagnetic indices calculated and made available by ISGI Collaborating Institutes from data collected at magnetic observatories. We thank the involved national institutes, the INTERMAGNET network and ISGI (isgi.unistra.fr). The OMNI data were obtained from the GSFC/SPDF OMNIWeb interface (at https://omniweb.gsfc.nasa.gov). Sunspot data from the World Data Center SILSO, Royal Observatory of Belgium, Brussels.

G. Bernoux is thankful for funding from Région Occitanie and ONERA, under Grant Agreements 19008721/ALDOCT and 30196.

References

Akasofu, S.-I., 1981. Prediction of Development of Geomagnetic Storms Using the Solar Wind-Magnetosphere Energy Coupling Function ϵ. *Planetary and Space Science*, **29**(11), 1151–1158. 10.1016/0032-0633(81)90121-5. 2.1

Baker, D. N., E. W. Hones, J. B. Payne, and W. C. Feldman, 1981. A High Time Resolution Study of Interplanetary Parameter Correlations with AE. *Geophysical Research Letters*, **8**(2), 179–182. 10.1029/GL008i002p00179. 2.1

Baudin, M., A. Dutfoy, B. Iooss, and A.-L. Popelin, 2015. Open TURNS: An Industrial Software for Uncertainty Quantification in Simulation. *arXiv:1501.05242 [math, stat]*. 1501.05242. A

Berndt, D. J., and J. Clifford, 1994. Using Dynamic Time Warping to Find Patterns in Time Series. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, AAAIWS'94, 359–370. AAAI Press, Seattle, WA. 3.4.2

Bernoux, G., and V. Maget, 2020. Characterizing Extreme Geomagnetic Storms Using Extreme Value Analysis: A Discussion on the Representativeness of Short Data Sets. *Space Weather*, **18**(6), e2020SW002,450. 10.1029/2020SW002450. 1, 2.1, 2.2, 2.3.1, 3.3

Borovsky, J. E., and Y. Y. Shprits, 2017. Is the Dst Index Sufficient to Define All Geospace Storms? *Journal of Geophysical Research: Space Physics*, **122**(11), 11,543–11,547. 10.1002/2017JA024679. 2.2

Borovsky, J. E., and K. Yakymenko, 2017. Systems Science of the Magnetosphere: Creating Indices of Substorm Activity, of the Substorm-Injected Electron Population, and of the Electron Radiation Belt. *Journal of Geophysical Research: Space Physics*, **122**(10), 10,012–10,035. 10.1002/2017JA024250. 2.2

Burton, R. K., R. L. McPherron, and C. T. Russell, 1975. An Empirical Relationship between Interplanetary Conditions and Dst. *Journal of Geophysical Research (1896-1977)*, **80**(31), 4204–4214. 10.1029/JA080i031p04204. 2.1

Camporeale, E., 2019. The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. *Space Weather*, **17**(8), 1166–1207. 10.1029/2018SW002061. 1, 4.3

Carè, A., and E. Camporeale, 2018. Chapter 4 - Regression. In E. Camporeale, S. Wing, and J. R. Johnson, eds., Machine Learning Techniques for Space Weather, 71–112. Elsevier. ISBN 978-0-12-811788-0. 10.1016/B978-0-12-811788-0.00004-4. 2.3.1

Chakraborty, S., and S. K. Morley, 2020. Probabilistic Prediction of Geomagnetic Storms and the Kp Index. *Journal of Space Weather and Space Climate*, **10**, 36. 10.1051/swsc/2020037. 1, 2.1, 3.1, 5

Chandorkar, M., E. Camporeale, and S. Wing, 2017. Probabilistic Forecasting of the Disturbance Storm Time Index: An Autoregressive Gaussian Process Approach. *Space Weather*, **15**(8), 1004–1019. 10.1002/2017SW001627. 2.1, 5

Chi, Y., C. Shen, B. Luo, Y. Wang, and M. Xu, 2018. Geoeffectiveness of Stream Interaction Regions From 1995 to 2016. *Space Weather*, **16**(12), 1960–1971. 10.1029/2018SW001894. 2.3.1

Chi, Y., C. Shen, Y. Wang, M. Xu, P. Ye, and S. Wang, 2016. Statistical Study of the Interplanetary Coronal Mass Ejections from 1995 to 2015. *Solar Physics*, **291**(8), 2419–2439. 10.1007/s11207-016-0971-5. 2.3.1

Forsyth, C., C. E. J. Watt, M. K. Mooney, I. J. Rae, S. D. Walton, and R. B. Horne, 2020. Forecasting GOES 15 >2 MeV Electron Fluxes From Solar Wind Data and Geomagnetic Indices. *Space Weather*, **18**(8), e2019SW002,416. 10.1029/2019SW002416. 4.3

Frías-Paredes, L., F. Mallor, T. León, and M. Gastón-Romeo, 2016. Introducing the Temporal Distortion Index to Perform a Bidimensional Analysis of Renewable Energy Forecast. *Energy*, **94**, 180–194. 10.1016/j.energy.2015.10.093. 3.4.2

Ghil, M., M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, et al., 2002. Advanced Spectral Methods for Climatic Time Series. *Reviews of Geophysics*, **40**(1), 3–1–3–41. 10.1029/2000RG000092. 2.3.1, A

Gold, O., and M. Sharir, 2018. Dynamic Time Warping and Geometric Edit Distance: Breaking the Quadratic Barrier. *ACM Transactions on Algorithms*, **14**(4), 50:1–50:17. 10.1145/3230734. 3.4.2

Goodfellow, I., Y. Bengio, and A. Courville, 2016. Deep Learning. The MIT Press, Cambridge, Massachusetts, illustrated edition edn. ISBN 978-0-262-03561-3. 3.1

Gruet, M., M. Chandorkar, A. Sicard, and E. Camporeale, 2018. Multiple-Hour-Ahead Forecast of the Dst Index Using a Combination of Long Short-Term Memory Neural Network and Gaussian Process. *Space Weather*, **16**(11), 1882–1896. 10.1029/2018SW001898. 1, 3.1, 3.4.1, 4.3

Guen, V. L., and N. Thome, 2019. Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models. *arXiv:1909.09020 [cs, stat]*. 1909.09020. 3.4.2

Hochreiter, S., 1998. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **06**(02), 107–116. 10.1142/S0218488598000094. 3.1

Hochreiter, S., and J. Schmidhuber, 1997. Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780. 10.1162/neco.1997.9.8.1735. 3.1

Horne, R. B., S. A. Glauert, N. P. Meredith, D. Boscher, V. Maget, D. Heynderickx, and D. Pitchford, 2013. Space Weather Impacts on Satellites and Forecasting the Earth's Electron Radiation Belts with SPACECAST. *Space Weather*, **11**(4), 169–186. 10.1002/swe.20023. 1

King, J. H., and N. E. Papitashvili, 2005. Solar Wind Spatial Scales in and Comparisons of Hourly Wind and ACE Plasma and Magnetic Field Data. *Journal of Geophysical Research: Space Physics*, **110**(A2). 10.1029/2004JA010649. 2.1

Kingma, D. P., and J. Ba, 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. 1412.6980. 3.2

Kondrashov, D., R. Denton, Y. Y. Shprits, and H. J. Singer, 2014. Reconstruction of Gaps in the Past History of Solar Wind Parameters. *Geophysical Research Letters*, **41**(8), 2702–2707. 10.1002/2014GL059741. 2.3.1, A

Kondrashov, D., and M. Ghil, 2006. Spatio-Temporal Filling of Missing Points in Geophysical Data Sets. *Nonlinear Processes in Geophysics*, **13**(2), n/a. 10.5194/npg-13-151-2006. A

Kondrashov, D., Y. Shprits, and M. Ghil, 2010. Gap Filling of Solar Wind Data by Singular Spectrum Analysis. *Geophysical Research Letters*, **37**(15). 10.1029/2010GL044138. 2.3.1, A

Laperre, B., J. Amaya, and G. Lapenta, 2020. Dynamic Time Warping as a New Evaluation for Dst Forecast With Machine Learning. *Frontiers in Astronomy and Space Sciences*, **7**. 10.3389/fspas.2020.00039, 2006.04667. 3.1, 3.4.2

Lazzús, J. A., P. Vega, P. Rojas, and I. Salfate, 2017. Forecasting the Dst Index Using a Swarm-Optimized Neural Network. *Space Weather*, **15**(8), 1068–1089. 10.1002/2017SW001608. 2.3.1, 3.4.1, 4.3

Lethy, A., M. A. El-Eraki, A. Samy, and H. A. Deebes, 2018. Prediction of the Dst Index and Analysis of Its Dependence on Solar Wind Parameters Using Neural Network. *Space Weather*, **16**(9), 1277–1290. 10.1029/2018SW001863. 1

Ling, A. G., G. P. Ginet, R. V. Hilmer, and K. L. Perry, 2010. A Neural Network–Based Geosynchronous Relativistic Electron Flux Forecasting Model. *Space Weather*, **8**(9). 10.1029/2010SW000576. 1

Lundstedt, H., and P. Wintoft, 1994. Prediction of Geomagnetic Storms from Solar Wind Data with the Use of a Neural Network. *Annales Geophysicae*, **12**(1), 19–24. 10.1007/s00585-994-0019-2. 2.1

Matéo-Vélez, J.-C., A. Sicard, D. Payan, N. Ganushkina, N. P. Meredith, and I. Sillanpäa, 2018. Spacecraft Surface Charging Induced by Severe Environments at Geosynchronous Orbit. *Space Weather*, **16**(1), 89–106. 10.1002/2017SW001689. 3.3

Mayaud, P.-N., 1971. Une Mesure Planétaire d'activité Magnétique Basée Sur Deux Observatoires Antipodaux. *Annales Geophysicae*, **27**, 67–70. 2.1

Mayaud, P.-N., 1980. Derivation, Meaning, and Use of Geomagnetic Indices. Geophysical Monograph ; 22. American Geophysical Union, Washington. ISBN 978-0-87590-022-3. 2.1

McComas, D. J., S. J. Bame, B. L. Barraclough, J. R. Donart, R. C. Elphic, J. T. Gosling, M. B. Moldwin, K. R. Moore, and M. F. Thomsen, 1993. Magnetospheric Plasma Analyzer: Initial Three-Spacecraft Observations from Geosynchronous Orbit. *Journal of Geophysical Research: Space Physics*, **98**(A8), 13,453–13,465. 10.1029/93JA00726. 3.3

Meredith, N. P., R. B. Horne, S. A. Glauert, R. M. Thorne, D. Summers, J. M. Albert, and R. R. Anderson, 2006. Energetic Outer Zone Electron Loss Timescales during Low Geomagnetic Activity. *Journal of Geophysical Research: Space Physics*, **111**(A5). 10.1029/2005JA011516. 2.1

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, et al., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, eds., Advances in Neural Information Processing Systems 32, 8024–8035. Curran Associates, Inc. 3.2

Riley, P., D. Baker, Y. D. Liu, P. Verronen, H. Singer, and M. Güdel, 2017. Extreme Space Weather Events: From Cradle to Grave. *Space Science Reviews*, **214**(1), 21. 10.1007/s11214-017-0456-3. 1

Rochel, S., D. Boscher, R. Benacquista, and J. F. Roussel, 2016. A Radiation Belt Disturbance Study from the Space Weather Point of View. *Acta Astronautica*, **128**, 650–656. 10.1016/j.actaastro.2016.07.012. 2.1, 2.2

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1986. Learning Representations by Back-Propagating Errors. *Nature*, **323**(6088), 533–536. 10.1038/323533a0. 3.2

Sexton, E. S., K. Nykyri, and X. Ma, 2019. Kp Forecasting with a Recurrent Neural Network. *Journal of Space Weather and Space Climate*, **9**, A19. 10.1051/swsc/2019020. 3.4.1, 4.3

Sobol', I. M., 1967. On the Distribution of Points in a Cube and the Approximate Evaluation of Integrals. *USSR Computational Mathematics and Mathematical Physics*, **7**(4), 86–112. 10.1016/0041-5553(67)90144-9. A

Tan, Y., Q. Hu, Z. Wang, and Q. Zhong, 2018. Geomagnetic Index Kp Forecasting With LSTM. *Space Weather*, **16**(4), 406–416. 10.1002/2017SW001764. 1, 3.4.1, 4.3

Vallance, L., B. Charbonnier, N. Paul, S. Dubost, and P. Blanc, 2017. Towards a Standardized Procedure to Assess Solar Forecast Accuracy: A New Ramp and Time Alignment Metric. *Solar Energy*, **150**, 408–422. 10.1016/j.solener.2017.04.064. 3.4.2

Vautard, R., P. Yiou, and M. Ghil, 1992. Singular-Spectrum Analysis: A Toolkit for Short, Noisy Chaotic Signals. *Physica D: Nonlinear Phenomena*, **58**(1), 95–126. 10.1016/0167-2789(92)90103-T. 2.3.1, A

Wei, L., Q. Zhong, R. Lin, J. Wang, S. Liu, and Y. Cao, 2018. Quantitative Prediction of High-Energy Electron Integral Flux at Geostationary Orbit Based on Deep Learning. *Space Weather*, **16**(7), 903–916. 10.1029/2018SW001829. 1

Wing, S., J. R. Johnson, E. Camporeale, and G. D. Reeves, 2016. Information Theoretical Approach to Discovering Solar Wind Drivers of the Outer Radiation Belt. *Journal of Geophysical Research: Space Physics*, **121**(10), 9378–9399. 10.1002/2016JA022711. 2.1

Wing, S., J. R. Johnson, J. Jen, C.-I. Meng, D. G. Sibeck, K. Bechtold, J. Freeman, K. Costello, M. Balikhin, and K. Takahashi, 2005. Kp Forecast Models. *Journal of Geophysical Research: Space Physics*, **110**(A4). 10.1029/2004JA010500. 2.1

Wintoft, P., and M. Wik, 2018. Evaluation of Kp and Dst Predictions Using ACE and DSCOVR Solar Wind Data. *Space Weather*, **16**(12), 1972–1983. 10.1029/2018SW001994. 3.4.2

Wintoft, P., M. Wik, J. Matzka, and Y. Shprits, 2017. Forecasting Kp from Solar Wind Data: Input Parameter Study Using 3-Hour Averages and 3-Hour Range Values. *Journal of Space Weather and Space Climate*, **7**, A29. 10.1051/swsc/2017027. 1

Wu, J.-G., and H. Lundstedt, 1997. Geomagnetic Storm Predictions from Solar Wind Data with the Use of Dynamic Neural Networks. *Journal of Geophysical Research: Space Physics*, **102**(A7), 14,255–14,268. 10.1029/97JA00975. 2.1

Appendix A: Gap-filling with Singular Spectrum Analysis (SSA)

To fill the missing values in our dataset we followed the SSA gap-filling method described in Kondrashov et al. (2010) and very well summarised in Section 2 of Kondrashov et al. (2014): "SSA is a data-adaptive, nonparametric method for spectral estimation; a comprehensive review can be found in Ghil et al. (2002). It is based on diagonalization of the time-lagged covariance matrix of multivariate time series; the set of its eigenvectors or temporal empirical orthogonal functions (EOFs) is an optimal set of data-adaptive, narrowband filters for decomposing the variance

within a sliding time window M. Projecting the data set onto each EOF yields the corresponding principal component (PC); the entire time series or parts thereof can be reconstructed by using linear combinations of PCs and EOFs for selected number of K modes, which yield the reconstructed components. Kondrashov and Ghil (2006) developed an SSA-based gap-filling method that relied on the presence of significant oscillatory modes in the time series [...]. Kondrashov et al. (2010) generalized the SSA gap-filling methodology to multivariate geophysical data consisting of gappy "drivers" and continuous "response" records and applied it to fill in large gaps in solar wind and IMF data, by combining it with time-continuous geomagnetic indices. It is the covariation in driver and response at times when both are present that allows us to reconstruct the former when only the latter is measured."

Here we will be using the geomagnetic indices *Kp* and *Dst* as our "response" records. All the steps of the SSA gap-filling method are automatically performed by the SSA-MTM toolkit (Vautard et al., 1992) that is publicly available e.g. at `https://dept.atmos.ucla.edu/tcd/download`. But we also need to find the optimal M and K values, which are non-trivial. Kondrashov et al. (2014) suggests some values for a few solar wind parameters, but they did not include e.g. the plasma temperature T. That is why we performed a new search for optimal parameters, using a more recent period.

In order to find the optimal SSA window size M and number of modes K for each solar wind parameter, we introduced artificial gaps in each time series for the period 2008-2018. The artificial gaps are reproductions of the true gaps found in the same time series, but during the period 1984-1994, so that the distribution and the length of the artificial gaps were plausible. Then we searched for the best M and K values that allowed for the best reconstruction of the gaps as measured with the RMSE and Pearson correlation. The parameters optimal parameters were found by performing an iterative grid search using quasi-random low discrepancy Sobol sequences (Sobol', 1967) generated by the Python package OpenTURNS (Baudin et al., 2015). The set of optimal parameters (namely M* and K*) found for each time series are reported in Table A.1. The time series gap-filled using the SSA technique and these M* and K* values are available online as supplementary material to this article.

Table A.1. Optimal M* and K* values found for filling the gaps in the time series.

Solar wind parameter	M*	K*
V_{sw}	110	29
ρ	12	30
T	12	29
B_z	9	17