Geophysical Journal International

Geophys. J. Int. (2022) **231,** 520–535
Advance Access publication 2022 May 30
GJI Geomagnetism, Rock Magnetism and Palaeomagnetism

https://doi.org/10.1093/gji/ggac195

Can machine learning reveal precursors of reversals of the geomagnetic axial dipole field?

K. Gwirtz,[1] T. Davis,[2] M. Morzfeld,[2] C. Constable ,[2] A. Fournier [3] and G. Hulot[3]

[1] *NASA Postdoctoral Program Fellow, NASA Goddard Space Flight Center, Greenbelt, MD* 20771, *USA. E-mail: kylegwirtz@gmail.com*
[2] *Cecil H. and Ida M. Green Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego, CA 92093-0225, USA*
[3] *Université Paris Cité, Institut de Physique du Globe de Paris, CNRS, F-*75005 *Paris, France*

Accepted 2022 May 18. Received 2022 May 18; in original form 2022 March 7

SUMMARY

It is well known that the axial dipole part of Earth's magnetic field reverses polarity, so that the magnetic North Pole becomes the South Pole and vice versa. The timing of reversals is well documented for the past 160 Myr, but the conditions that lead to a reversal are still not well understood. It is not known if there are reliable 'precursors' of reversals (events that indicate that a reversal is upcoming) or what they might be. We investigate if machine learning (ML) techniques can reliably identify precursors of reversals based on time-series of the axial magnetic dipole field. The basic idea is to train a classifier using segments of time-series of the axial magnetic dipole. This training step requires modification of standard ML techniques to account for the fact that we are interested in rare events—a reversal is unusual, while a non-reversing field is the norm. Without our tweak, the ML classifiers lead to useless predictions. Perhaps even more importantly, the usable observational record is limited to 0–2 Ma and contains only five reversals, necessitating that we determine if the data are even sufficient to reliably train and validate an ML algorithm. To answer these questions we use several ML classifiers (linear/non-linear support vector machines and long short-term memory networks), invoke a hierarchy of numerical models (from simplified models to 3-D geodynamo simulations), and two palaeomagnetic reconstructions (PADM2M and Sint-2000). The performance of the ML classifiers varies across the models and the observational record and we provide evidence that this is *not* an artefact of the numerics, but rather reflects how 'predictable' a model or observational record is. Studying *models* of Earth's magnetic field via ML classifiers thus can help with identifying shortcomings or advantages of the various models. For Earth's magnetic field, we conclude that the ability of ML to identify precursors of reversals is limited, largely due to the small amount and low frequency resolution of data, which makes training and subsequent validation nearly impossible. Put simply: the ML techniques we tried are not currently capable of reliably identifying an axial dipole moment (ADM) precursor for geomagnetic reversals. This does not necessarily imply that such a precursor does not exist, and improvements in temporal resolution and length of ADM records may well offer better prospects in the future.

Key words: Dynamo: theories and simulations; Magnetic field variations through time; Palaeointensity; Reversals: process, time scale, magnetostratigraphy; Time-series analysis.

1 INTRODUCTION

Computers and hand-held devices have become a normal part of our daily lives and along with computers came the broad use of statistical algorithms, typically referred to as machine learning (ML) or artificial intelligence (AI). By now, ML and AI are encountered daily: the algorithms sort our email for spam, suggest the next video we want to watch, assist in completing our tax returns, and present us with advertisements that are deemed of interest. The incredible success of ML/AI is in large part due to the availability of massive amounts of data: looking through vast amounts of emails makes it possible to identify features that render an email suspicious. Another reason for the success of ML/AI is that very simple strategies can often be very successful: it is likely that you will enjoy watching a video very similar to the one you just enjoyed watching. Simple strategies are easy to discover. Finally, if the ML/AI algorithm makes

a mistake, the consequences are usually 'minor'—the company makes less money because the advertisement strategy is suboptimal, or you may need to delete a few or a lot of additional emails.

None of the above is generally true in Earth science or in geophysics and the study of Earth's deep interior. There are no vast amounts of data—every measurement and observation is the result of a long, costly effort. Simple prediction strategies are useless—I may often turn out to be right when I predict that the weather tomorrow will be the same as the weather today, but such a prediction strategy misses the point of predicting *changes* in the current conditions. And finally, a 'wrong' assessment or prediction can have disastrous consequences, for example when predicting the path of a hurricane.

Nonetheless, there are many ingenious and careful efforts to port the success of ML and AI into Earth science, keeping the above mentioned problems in mind. We follow this path as well. The problem we are concerned with is predicting reversals in the polarity of the Earth's axial magnetic dipole field. Such reversals have occurred numerous times throughout Earth's history (Cande & Kent 1995; Lowrie & Kent 2004; Ogg 2012), most recently around 780 kyr ago (the Brunhes–Matuyama reversal). And while the occurrence and timing of reversals is well documented, the conditions that lead to a reversal are not fully understood. We note that while studies of simulations suggest that detailed predictions of the geomagnetic field may be limited to less than a century (Hulot *et al.* 2010; Lhuillier *et al.* 2011a), 'coarse' predictions of macroscopic features of the field may be possible over much longer timescales (Morzfeld *et al.* 2017). Indeed this possibility has led to multiple studies aimed at searching for precursors—events or patterns of behaviour that indicate an upcoming reversal. For example Olson *et al.* (2009) investigate a dynamo model during two periods of dipole collapse and highlight, for example, patterns of reverse flux patches as potential precursors. Other examples include (e.g. Constable & Korte 2006; Laj & Kissel 2015; Valet & Fournier 2016; Brown *et al.* 2018), which carefully study the past behaviour Earth's magnetic field leading up to reversals, in particular with regards to the (fast) decay in intensity of the modern field.

A natural idea is to use ML to search for precursors to reversals within the time evolution of Earth's magnetic field. Here, we are limited to searching for precursors in reconstructed time-series of the axial dipole (or virtual axial dipole moment) of Earth's magnetic field, because these are the only 'data' available. The non-dipole field is not well documented over the geological timescales of millions of years, relevant to the dynamics of reversals.

We search for precursors of reversals of Earth's magnetic field using 'classifiers' (see, e.g. Goodfellow *et al.* 2016). The basic idea of a classifier is simple. One can train ML algorithms to sort input data into two (or more) classes. The two classes can, for example, be 'cats' and 'dogs'. The procedure is to feed a large set of 'training data' to an ML algorithm and to subsequently validate the algorithm on independent 'validation data' to avoid overfitting. The training data are a large collection of images of cats and dogs, each image being accompanied by a 'label', indicating whether the image contains a dog or a cat, while the validation data consist of a *different* set of labelled images of cats and dogs. The validation step is crucial to avoid overfitting. After training and validation, the algorithm can be used to classify new images. To port these ideas to reversals of Earth's magnetic field we swap 'cats and dogs' for segments of a time-series that either precede a reversal event or not. The rest of this paper is dedicated to determining under what circumstances this simple idea might actually be useful.

A first difficulty is caused by the fact that reversals are rare (five reversals over the past 2 Myr), which means that the data are *imbalanced*. The difficulty of training standard ML algorithms with imbalanced data is that they tend to favour strategies which may learn to assign a single output to every input. For example, when the algorithm, during training, almost exclusively encounters images of cats, the ML may 'think' that every image is that of a cat. This is due to the fact that the ML often optimizes 'accuracy' (or similar loss functions) during training. Accuracy is the ratio of correct classifications to the number of classifications made. If the data are imbalanced, one can achieve a high accuracy via useless prediction strategies, that assign the same output to every input (see also Gwirtz *et al.* 2021). To address this issue we tweak standard ML techniques, specifically linear and non-linear support vector machines (SVM) and long short-term memory networks (LSTM, see, e.g. Hochreiter & Schmidhuber 1997; Cristianini & Shawe-Taylor 2000), to penalize false negatives (failing to correctly predict a reversal) more severely than false positives (incorrectly predicting that a reversal will occur) during the training period. Our tweak is an effective, yet computationally intensive way to deal with imbalanced data and may prove useful in other applications of ML.

A second difficulty, which is harder to overcome, is the limited amount of data we can use to train and validate an ML algorithm. The observational record of Earth's axial magnetic dipole field is limited and the palaeomagnetic reconstructions we consider cover just the last 2 Myr, containing five reversals (PADM2M (Ziegler *et al.* 2011) and Sint-2000 (Valet *et al.* 2005)). This may severely limit what any ML algorithm can do. To address the limited amount of data, we study ML also in the context of model output from computational simulations because these 'data' are not limited. The goal is to use models to discover what is in principle feasible (large training and/or validation data sets), to better interpret results obtained in practice (limited data). The models we consider are the simplified differential equation model of Gissinger (2012), a reversing 3-D dynamo simulation and the stochastic models of Pétrélis *et al.* (2009) and Morzfeld & Buffett (2019).

We note that a reversal must *always* follow a period during which the axial dipole field is weak. The reason is the continuity of the evolution of Earth's magnetic field (which is undisputed). The strength of the axial dipole must drop to a low value, before it can ultimately collapse to zero, and then re-build in opposite polarity. Thus, a simple precursor of a reversal is an intensity threshold. If the intensity of the axial dipole field drops below the threshold, a reversal may be unavoidable (or at least very likely to occur). This idea has been studied in detail in Gwirtz *et al.* (2021), clearly spelling out the capability and limitations of threshold-based predictions of Earth's magnetic reversals. The present paper can be viewed as an extension of this previous work, looking in detail into the question of whether there are *dynamics* in the axial dipole field that indicate an upcoming reversal. For that reason, we benchmark the ML techniques of this study against the simpler threshold-based predictions of Gwirtz *et al.* (2021). ML is only useful if it can reliably outperform simpler methods. Further illumination of our experiments with ML is provided by analysis of the autocorrelation structure of our various models and palaeomagnetic reconstructions that allow us to better document the evolutionary nature of the reversals and what makes them more or less predictable.

The remainder of this paper is organized in the following way. In Section 2 we introduce the models and palaeomagnetic reconstructions that we use, provide relevant background on threshold-based predictions, and briefly outline SVMs and LSTMs. In Section 3, we describe the procedures we use to train and validate ML algorithms

for use in predicting reversals of Earth's axial dipole field. The results of a collection of numerical experiments are presented in Section 4 followed by a discussion of their robustness and geophysical significance in Section 5. We present conclusions in section 6.

2 BACKGROUND: MODELS, PALAEOMAGNETIC RECONSTRUCTIONS, THRESHOLD-BASED PREDICTIONS AND MACHINE LEARNING

We summarize relevant background materials, beginning with a brief review of numerical models for Earth's axial dipole field and palaeomagnetic reconstructions of that field. Next, we review previous work on threshold-based predictions, and the definitions of low-dipole events and prediction horizons. Finally, we describe the ML methods we use in this paper.

2.1 Numerical models and palaeomagnetic reconstructions

The low-dimensional model we use is that of Gissinger (2012) and consists of the coupled ordinary differential equations

$$\begin{aligned}\frac{\mathrm{d}Q}{\mathrm{d}t} &= \mu Q - V D,\\ \frac{\mathrm{d}D}{\mathrm{d}t} &= -\nu D + V Q,\\ \frac{\mathrm{d}V}{\mathrm{d}t} &= \Gamma - V + Q D,\end{aligned} \tag{1}$$

where $\mu = 0.119$, $\nu = 0.1$, and $\Gamma = 0.9$. The three scalar values of Q, D and V are representative of the quadrupole, dipole and fluid velocity, respectively. The sign of D indicates polarity (today's or reversed polarity) so that a change in sign corresponds to a dipole reversal. We refer to this model as the G12 model and use the G12 millennium timescale (1 dimensionless time unit = 4 kyr) of Morzfeld *et al.* (2017) to convert model time into geological time. A segment of the dipole of a G12 simulation is shown in the top panel of Fig. 1. We note that we only work with the D variable of the G12 model, which serves as a proxy for the Earth's axial dipole. The other two variables of G12 remain opaque to all ML/thresholding methods we use. The reason is that only the dipole field of Earth is observable. We briefly bring up other simplified models in the context of some numerical experiments, but do not describe these in detail here (see Gwirtz *et al.* 2021, for more details on other simplified models).

We additionally consider a 3-D numerical dynamo simulation which exhibits polarity reversals. The simulation is part of an ensemble of reversing simulations run by N. Schaeffer (ISTerre, CNRS, Université Grenoble Alpes), A. Fournier and T. Gastine (both affiliated with Université Paris Cité, Institut de Physique du Globe de Paris). The time-series of the axial dipole from this simulation was previously studied in Gwirtz *et al.* (2021), where further details of the numerical model and its favourable comparisons with the geomagnetic field can be found. Time in the non-dimensional simulation is scaled such that the secular-variation timescale matches that of Earth (415 yr, see Lhuillier *et al.* 2011b). A segment of the time-series of axial dipole intensity from this simulation can be seen in the bottom panel of Fig. 1. The entire simulation covers 147 Myr (at a time step of 43.09 years) and contains about 360 low-dipole events (see below for our explicit definition), 109 of which are reversals.

We also consider the PADM2M and Sint-2000 palaeomagnetic axial dipole moment (PADM) reconstructions (Valet *et al.* 2005; Ziegler *et al.* 2011), derived from estimates of the virtual axial dipole moment (VADM) of Earth's field over the last 2 Myr. The geomagnetic polarity timescale of Cande & Kent (1995) is used to determine the timing of reversals with a modification made in PADM2M for the Cobb mountain sub-chron (see Morzfeld *et al.* 2017). Both reconstructions are evaluated at time steps of 1 kyr and are shown in Fig. 2, where, as in G12, the sign indicates polarity and we accordingly have introduced a change in sign corresponding to reversals of the axial dipole field. Here, the PADMs are scaled so that their time average is equal to one. We note that despite covering the same period of time there are some differences between PADM2M and Sint-2000 resulting from the selection, volume, and interpretation of available data. This reflects an inherent uncertainty in reconstructing the past magnetic field of the Earth which we must be aware of when working with and drawing conclusions from palaeomagnetic reconstructions (Morzfeld *et al.* 2017).

2.2 Review of threshold-based reversal predictions

The ML techniques we present here extend the framework of threshold-based predictions and we therefore provide a review of these ideas. For more details, we refer to Gwirtz *et al.* (2021).

2.2.1 Low-dipole events

The numerical experiments of Section 4 involve training SVMs to predict low-dipole events. We define low-dipole events as in Gwirtz *et al.* (2021). Specifically, a low-dipole event starts when the axial dipole changes sign or its intensity drops below a value called the start-of-event threshold (ST) and it ends when the intensity recovers to above a second value called the end-of-event threshold (ET). This definition is illustrated in Fig. 3 where ST and ET are indicated by light blue and green horizontal lines, respectively, and a low-dipole event is highlighted in blue. This approach allows spans of time during which intensity is low to be identified as one single event. For example, the Cobb mountain subchron, seen in PADM2M and Sint-2000 around 1.2 Myr in the past (Fig. 2) is a single low-dipole event (and not a sequence of two reversals). To make this definition comparable across the models and reconstructions, axial dipole time-series are scaled to the respective model or reconstruction's average intensity and thresholds are expressed as a percentage of that average. Following Gwirtz *et al.* (2021), we use ST=10 per cent and ET=80 per cent, which means that an event starts when the intensity drops below 10 per cent of the average value and it ends when it exceeds 80 per cent of the average value.

2.2.2 Prediction horizons

The reversal predictions we make are 'coarse' in the sense that we predict whether a low-dipole event will begin within an *a priori* specified period of time, called the prediction horizon (PH)—the precise timing of the reversal is not considered (Morzfeld *et al.* 2017). This is illustrated in Fig. 3 where the length of the double-sided purple arrow indicates the PH. In the illustration, an event does occur in the future, but beyond the PH. Therefore, the correct prediction would be that no low-dipole event begins within the PH.

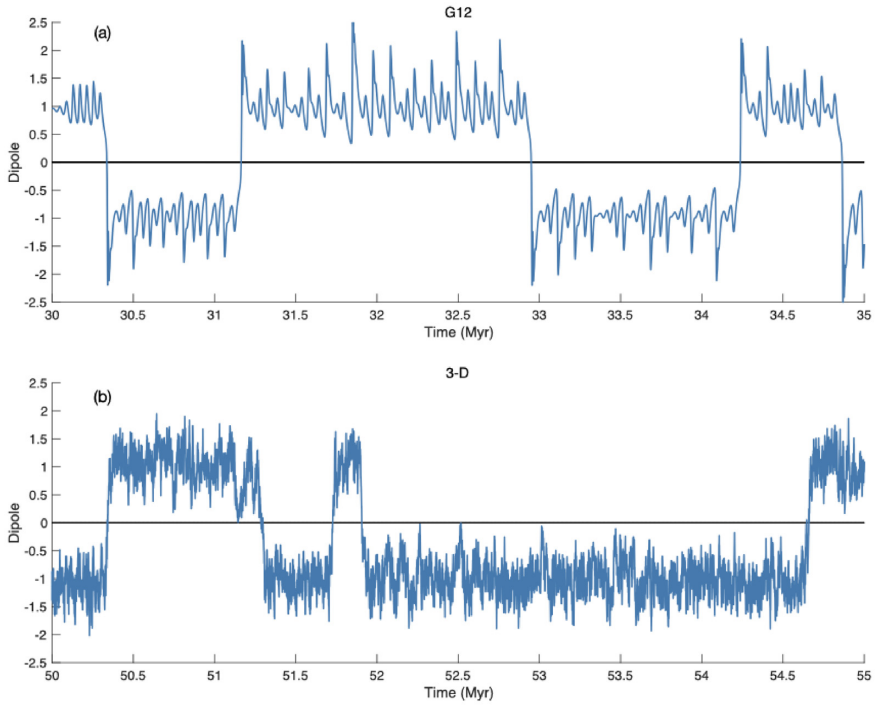

Figure 1. Axial dipole intensity as a function of time for the (a) G12 and (b) 3-D models (see Section 2.1). The sign indicates polarity and the amplitude is scaled such that the average intensity is one. The time scaling of both models is explained in the text.

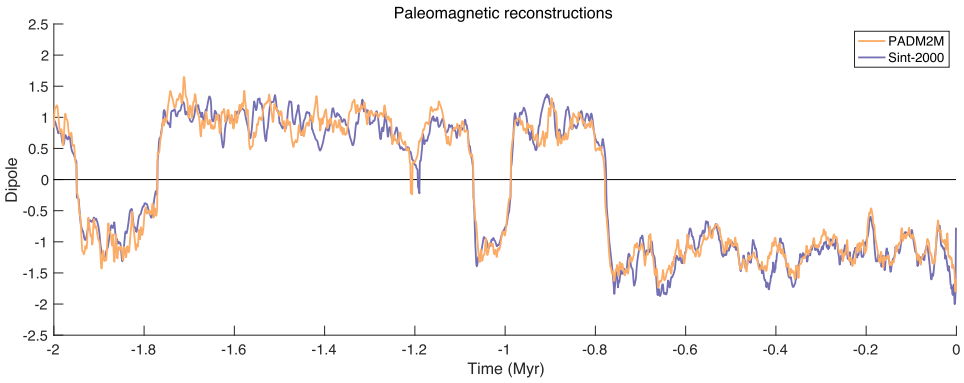

Figure 2. Modified versions of the PADM2M (yellow) and Sint-2000 (purple) palaeomagnetic reconstructions of the axial dipole (PADM) covering the last 2 Myr. The sign indicates polarity and the amplitude is scaled such that the average intensity for each time-series is one.

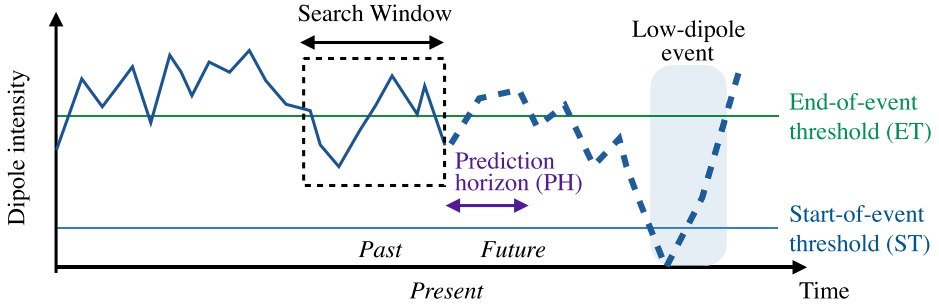

Figure 3. Illustration of the ML prediction strategy. The thick blue line represents a time-series of axial dipole intensity. The thin blue and green horizontal lines show the start-of-event and end-of-event thresholds (which define low-dipole events). The search window (dashed rectangle) encompasses the time-series segment used to make a prediction of whether a low-dipole event will start within a prediction horizon (purple arrow) from the present (to the right of the search window).

Table 1. Prediction horizons (equal to average event durations), in kyr, of the models and palaeomagnetic reconstructions.

G12	3-D	PADM2M	Sint-2000
3.2	16.4	11.7	10.2

The PH needs to be chosen prior to any training and validation and should be such that one anticipates the start of a low-dipole event by a meaningful span of time. Specifically, if the PH is short, we may only predict events when they have already begun. Conversely, for an excessively long PH, anticipating an event becomes trivial because reversals/low-dipole events are likely to occur within a very large time window. In Gwirtz *et al.* (2021), time was rescaled such that one dimensionless time unit was equal to the average duration of a low-dipole event in each respective model and reconstruction. Prediction horizons of one dimensionless time unit were then used and it was verified that subsequent results are not sensitive to the choice of PH (within the range of a few event durations). Here, we also use a PH of one average event duration, listed in Table 1 for the G12 and 3-D models and the two palaeomagnetic reconstructions.

2.2.3 Threshold-based predictions

The threshold-based prediction strategy of Gwirtz *et al.* (2021) is as follows: predict that a low-dipole event will begin within the prediction horizon whenever the axial dipole intensity is below a predefined level called the *warning threshold*. During training, we can label the true outcome as either positive (P), a low-dipole event begins within the prediction horizon or, negative (N), an event does not begin within the prediction horizon. Each prediction of the threshold-based strategy can thus be labelled as a true positive (TP), true negative (TN), false positive (FP) or false negative (FN). An 'optimal' warning threshold is determined during the training phase by applying a collection of candidate warning thresholds to a set of training data, tabulating the total number of true/false positives/negatives, and selecting the warning threshold which maximizes the Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{2}$$

A perfect MCC score is one, while a score near zero (or an undefined score in the case where the denominator is zero) indicates poor predictions (and negative MCC scores suggest to reverse the prediction strategy altogether). The MCC is useful as it is robust when evaluating classifications of imbalanced data (Chicco & Jurman 2020) like those we face in the prediction of low-dipole events. Specifically, low-dipole events are rare and therefore one could obtain a high *accuracy*, which is another popular skill score, by always predicting that no event will occur. To see why, recall that accuracy is defined as the ratio of all correct predictions (TP and TN), and all predictions made (P+N):

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}. \tag{3}$$

If $P \ll N$ (very few low-dipole events), then ACC can be large, even if TP is zero (we never predict that an event will occur). As a specific example, consider the case of training data with only one low-dipole event out of 100 training instances (P = 1, N = 99). A strategy that assigns N (no low-dipole event) to any input scores an accuracy of ACC = 99 per cent (because TP=0, TN = 99). This strategy, however, cannot be useful because it misses the point of being able to predict that a low-dipole event can occur. The use of the MCC avoids these issues—indeed, the MCC for a strategy which always makes the same prediction, as in the above example, is undefined (poor predictions). Further details of threshold-based predictions and their skills are discussed in detail in (Gwirtz *et al.* 2021). In this paper, threshold-based predictions serve as a baseline for the performance of the more sophisticated ML techniques.

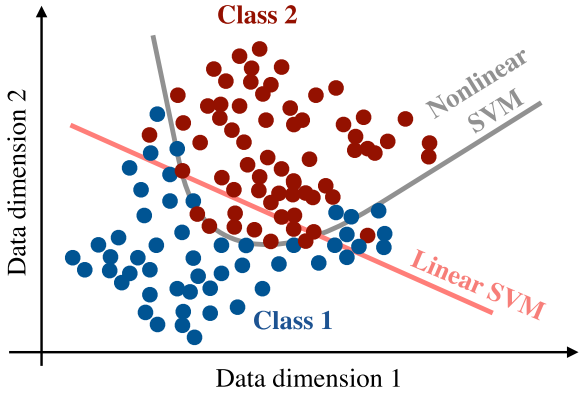

Figure 4. Illustration of SVMs. 2-D training data (dots) are labelled according to their classes, with class 1 being 'blue' and class 2 being 'red.' A linear SVM attempts to separate the two classes by a hyperplane (orange line), obtained via optimization of a loss function. A non-linear SVM effectively lifts the training data into a higher dimensional space where a separating plane is determined, again via optimization of a loss function. After projecting back into the original space, the classifier of the non-linear SVM—which is a plane in the higher dimensional space—appears as a curve (grey line) in the lower dimensional space (see text and Cortes & Vapnik 1995,for details).

2.3 Machine learning methods

We primarily use linear SVMs to search for precursors of reversals and major excursions (defined as events during which the field intensity drops below 10 per cent of its typical value, see above). This is done by training them to classify segments of time-series according to whether they precede low-dipole events (see Section 3). We use SVMs because they have been around for nearly three decades (Cortes & Vapnik 1995) and have since become a fundamental and widely used ML technique (see, e.g. Kim 2003; Ben-Hur *et al.* 2008; Ma & Guo 2014; Murty & Raghava 2016; Kok *et al.* 2021). Additionally, linear SVMs are conceptually easy to understand. In simple terms, training a linear SVM for binary classification—the case where there are only two possible classes—works as follows. Suppose the objects one wishes to classify are described by an n-dimensional vector and one has numerous examples of these objects, along with their correct classification (training data). The training of a linear SVM amounts to determining the $(n-1)$-dimensional hyperplane which separates the two classes with the maximum possible margin. In the case where the classes are not completely separable, training can include the minimization of a loss function which is dependent on the distance of misclassified objects from the candidate hyperplane (and thus their correct class).

Fig. 4 illustrates, via a cartoon, how SVMs separate training data. Specifically, the data are 2-D, and each dot in the $x-y$-plane is one datum that is labelled as a member of class 1 (blue) or class 2 (red). The orange line is the separating hyperplane (a line in this 2-D example) obtained by the SVM via optimizing the loss function. In this example, the classes are not linearly separable

and the linear SVM tries to settle on a line which achieves some separation while minimizing the extent of misclassifications (see appendix for further details). When the SVM is tasked to classify new data, it will assign classes according to which side of the separating hyperplane the new data fall on.

We also acknowledge that results could be sensitive to the particular choice of ML method. This can, again, be illustrated by the cartoon in Fig. 4, where the data are not linearly separable. A *non-linear* SVM effectively transforms data, which may not be linearly separable, to a higher-dimensional space where it is potentially linearly separable (see, e.g. Cristianini & Shawe-Taylor 2000). In the original space, this means that the boundaries that separate the two classes can be curved surfaces rather than simple hyperplanes. In the cartoon in Fig. 4, the result of a non-linear SVM is the grey curve, which arguably better separates the two classes.

For our case of searching for precursors of reversals, the above implies that the failure of a particular ML technique *cannot* be taken as conclusive evidence of the absence of such precursors. For this reason we include a number of experiments with techniques beyond linear SVMs, specifically with non-linear SVMs and with *long short-term memory* networks (LSTMs, Hochreiter & Schmidhuber 1997), which are recurrent neural networks often used for classifying time-series (see, e.g. Graves 2012). While we cannot exclude the possibility that other ML techniques may succeed at finding a precursor, the consistency of results we get with three different techniques allows us to be somewhat confident in our findings.

We rely on Matlab's machine learning toolbox to implement the (non)linear SVMs and the LSTMs. Specifically, we use `fitcsvm` to compute SVMs and `trainNetwork` (Matlab 2021a) to train LSTMs. Our code can be found on Github (https://github.com/kjg136/MLdipolePredictions). Due to the imbalanced nature of the data mentioned in Section 1, we make a small adjustment to the standard training algorithms. The details of these modifications are discussed in Section 3 and the Appendix.

Finally, we want to bring up the possibility to interpret ML methods within the (perhaps more familiar) framework of inverse problems: we select our model (SVMs) and search for parameters (a hyperplane) which optimize a misfit function (the extent of the separation of classes). Similar to an inverse problem, ML methods risk overfitting if the selected ML model has many free parameters, but not all of them are fully constrained by the (training) data. It is, however, difficult to interpret and account for errors in the data in ML methods (although this is trivial in inverse problems). Particularly in classifiers, labels are typically assumed error free—the image is either that of a cat or a dog, it cannot be an image of a cat-like dog or dog-like cat. Within classifiers, it thus remains unclear how to account for errors in the data and, for that reason, we do not explicitly consider uncertainties associated with the palaeomagnetic reconstructions. Rather, we interpret our results with the understanding that those uncertainties may have an impact on the outcomes of our experiments.

3 TRAINING SVMS WITH IMBALANCED DATA

The setup for making predictions using machine learning is as follows. At a given point in a time-series of axial dipole intensity, we search for precursors of low-dipole events by examining the recent past. Subsequently we refer to the time interval in which we search for precursors of low-dipole events as the *search window* (in kyr). After applying ML to the search window, we make a prediction of whether a low-dipole event will begin within the prediction horizon (PH, see Section 2.2.2). This is illustrated in Fig. 3 where the search window (dashed rectangle) contains the past segment of the time-series (thick blue line) being examined for low-dipole event precursors, and the length of the double-sided purple arrow indicates the size of the PH. We do not make predictions once an event has started, or if the search window overlaps with the end of an event. Each prediction is compared to the 'true' outcome, which is labelled either, Positive (P), or Negative (N), depending on whether a low-dipole event begins within the PH. In this way, the challenge of making predictions is a binary classification problem (Goodfellow *et al.* 2016).

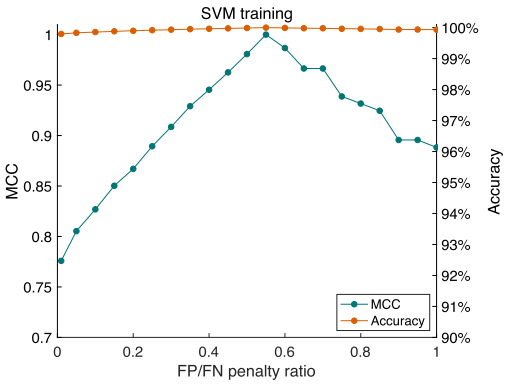

Figure 5. Accuracy (orange) and MCC (teal) as a function of the relative size of penalty applied to false positives compared to false negatives, during SVM training. The results shown here correspond to a classification problem of a G12 simulation (see text for more details).

To put it simply, during training we examine segments of an axial dipole time-series within the search window and train the SVM to determine whether it belongs to the class P or N, that is whether an event will soon begin or not. For example, at the time a prediction is being made in Fig. 3, the correct classification for the time-series segment in the search window is N; an event starts in the future but not within the prediction horizon. Thus, every prediction (classification) within the training data results in either a TP, TN, FP or FN. To avoid overfitting the training data, a time-series is divided into independent *training* and *validation* data. The training data is used in determining the SVM which is then applied to predictions with the validation data.

To avoid issues with imbalanced data, we use MCC during training of the SVM. But ML in general and SVM in particular are typically not designed to work with MCC, but rather to minimize a different 'loss function' (which we already indicated is problematic when data are imbalanced). The SVM code we use, for example, minimizes a loss function which measures the extent of separation of classes, and the extent of misclassifications by a hyperplane.

We tweak this training process to be more robust to imbalanced data in the following, non-intrusive way (non-intrusive meaning that we do not need to modify the actual SVM code). During training of an SVM, we successively reduce the weight given to misclassified negatives (false positives), relative to misclassified positives (false negatives) in the loss function (see the Appendix). In effect, this emphasizes the need to correctly classify the rare positives (P) over the more common negatives. We then select as optimal, the SVM which maximizes MCC on the training data. This process is illustrated in Fig. 5, where we train an SVM on a G12 simulation. Shown (in teal) is the MCC achieved during the training of an SVM

on a short simulation of G12, as a function of the relative size of penalty applied to false positives compared to false negatives (FP/FN penalty ratio). We note that MCC varies between 0.77 and nearly 1 with the FP/FN penalty ranging from zero to one. We select the SVM trained with the FP/FN penalty ratio of 0.55 as optimal because it achieves the highest MCC. For comparison, we also show accuracy (in orange) as a function of the FP/FN penalty, but accuracy remains relatively flat (and near 100 per cent), as would any strategy that only rarely (or never) predicts low-dipole events to occur.

Our tweak to use the MCC (rather than other loss functions) is non-intrusive and essentially amounts to embedding existing SVM code within a loop where the FP/FN penalty ratio is gradually reduced. We opted for this non-intrusive adjustment to be able to use several ML techniques, without having to write code for each one from scratch. Writing new ML code would entail computing gradients of the MCC cost function, which is computationally costly if finite differences are used, or conceptually difficult (and tedious to code) when the gradients are computed analytically or via automatic differentiation.

Finally, we repeat the entire training process for a number of search window lengths. In this way, we can assess the skill (MCC) of the classifier as a function of the time interval over which we search for precursors of reversals. If the search window is short, the skill of SVMs should be comparable to the skill of threshold-based predictions. If the search window contains only one point, the SVM is essentially finding a threshold (but the SVM threshold can differ from the threshold of Gwirtz *et al.* (2021) due to different numerical implementations of the threshold search). If the SVM skill increases with search window, then this indicates that the dynamics leading up to a reversal contain precursors of reversals that the SVM can detect.

4 RESULTS

We present the results of numerical experiments with SVMs applied to the models and palaeomagnetic reconstructions. In all experiments, we use a PH of one average event duration (see Table 1). We also report the skill of threshold-based predictions (Gwirtz *et al.* 2021), using the same training and validation data, to benchmark the SVMs against simpler methods.

4.1 SVMs applied to G12 and 3-D models

We begin with presenting results obtained with the numerical models. Because the 'data' are model outputs, they are not limited in number, which will help us interpret results obtained with limited palaeomagnetic data. We then study the consequences of data being limited.

4.1.1 SVMs trained on long time-series

We apply SVMs under the ideal circumstances where one has large training and validation data sets, each containing a large number of events. Specifically, the training and validation data for the G12 model and the 3-D model contain 180 events each (totalling 360 events for training and validation). For both models and each search window considered, the SVM with FP/FN penalty ratio that maximizes MCC on the training data is labelled as optimal and is subsequently applied to the validation data. The MCC scores on validation data as a function of the search window are shown in Fig. 6. The red dots correspond to scores with G12 and the green dots correspond to scores with the 3-D model. The dashed horizontal lines indicate the MCC of threshold-based predictions. We find that events in G12 are fairly predictable (MCC near one) via SVMs or thresholds, while the predictions for the 3-D model are more challenging (MCC closer to zero for both methods).

We note that for G12, the SVMs with search windows greater than 10 kyr perform consistently better than threshold-based predictions [see the magnified plot of panel (b) in Fig. 6]. This indicates that the SVMs are identifying features in the recent history of G12 time-series which indicate that a low-dipole event is about to occur within the PH. The fact that the MCC score stabilizes beyond search windows of around 11 kyr for the G12 model, suggests that the SVMs are not finding additional information about upcoming low-dipole events when looking more than 11 kyr into the past. This implies that a search window of around 11 kyr is 'optimal' for G12 in the sense that the SVMs which look further back in time gain no advantage. This stabilizing of the MCC for large windows is to be expected: The state of the axial dipole far in the past, should be unrelated to the present and future behaviour, due to the chaotic nature of the G12 model.

For the 3-D model, we note that the skill of the SVM (green dots) is lower than the skill of the threshold-based technique (dashed green line) for every search window we tried. This is particularly surprising for a search window of 1 kyr, for which the SVM should indeed determine a threshold (a hyperplane in one dimension is a point). The differences we observe here are due to the fact that the threshold-based predictions *directly* search for a threshold that maximizes MCC, while the SVM we employ tweaks the established SVM machinery to find a threshold that maximizes MCC (see Section 3). With a finer grid of FP/FN penalty ratios the SVM skill can be brought closer to that of the threshold strategy, but we do not pursue this further here.

As we will explain in detail below, the interesting quantity is the *change* of skill with the size of search window, not the 'raw' MCC skill score. To that extent, we note that skill scores of the 3-D model do not change with the search window, while those of G12 do increase with search window size (up to a limit due to chaos). This means that the SVMs cannot detect precursors for reversals (beyond a threshold) in the 3-D model, but in the G12 model the recent past leading up to a low-dipole event indeed contains additional information (a precursor).

To get a qualitative understanding of what patterns the linear SVMs might find (or not find in the case of the 3-D simulation) we directly examine the lead-up to events in the G12 and 3-D time-series. Fig. 7 shows the axial dipole intensity 50 kyr ahead of a low-dipole event, that is at time $t = 0$, the axial dipole has dropped below 0.1 (10 per cent of the average intensity). The time-series segments shown are semi-transparent allowing the boldness of the colour to indicate the amount of overlap. From these plots, the situation is clear. The G12 model exhibits a consistent pattern that even the human eye can pick up, especially for shorter search windows (about 10 kyr). For the 3-D model, identifying a pattern by the human eye is difficult, even if the search window is relatively short ($<$10 kyr). Thus, one may argue that artificial intelligence/machine learning, does not increase the ability of the human eye significantly—perhaps indicating that a pattern, or precursor, exists for G12, but not for the 3-D model.

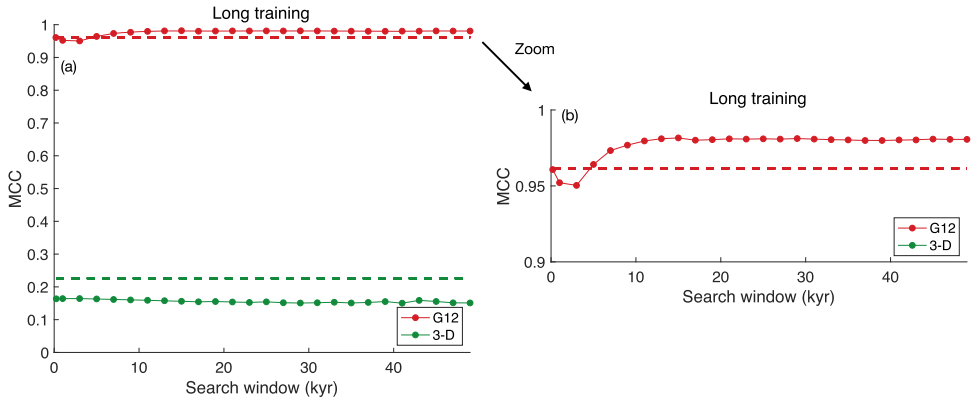

Figure 6. Validation MCC as a function of search window for SVMs trained on large data sets of the G12 (red) and 3-D (green) models. Horizontal dashed lines indicate the MCC of threshold-based predictions using the same training and validation data from G12 (red) and 3-D (green). Panel (a) shows the plot for MCC scores in the range 0–1. Panel (b) shows the same plot zoomed into the G12 results (MCC score range 0.9–1).

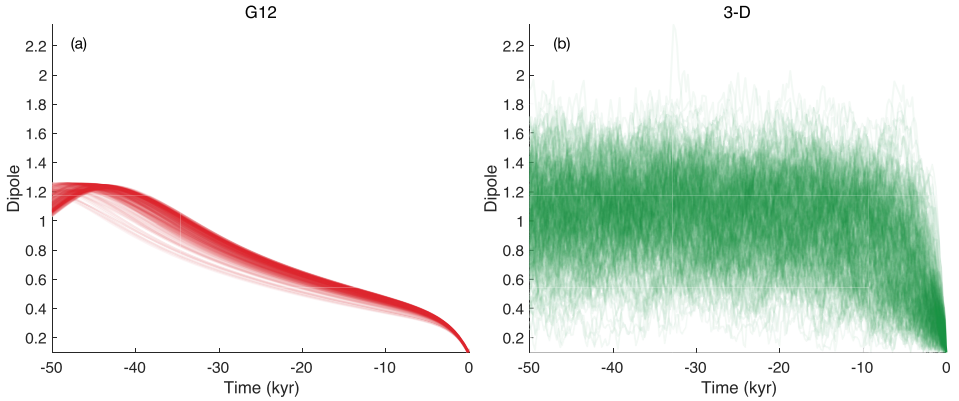

Figure 7. Time-series of 50 kyr of axial dipole intensity preceding the beginning of events for (a) G12 and (b) the 3-D model. The plotted intensities are semi-transparent with the amount of overlap indicated by the boldness of colour.

4.1.2 SVMs trained on short time-series

We note that while the time-series used to train and validate the SVMs in Fig. 6 are long, the observational record of the Earth's magnetic field is relatively short. The palaeomagnetic reconstructions of PADM2M and Sint-2000 (see Section 4.2) span just the last 2 Myr and contain only six low-dipole events. For this reason, we repeat the process of training and validation but this time, use training data containing only five events (the validation data, critically, remains the same). The resulting validation MCC is shown as a function of search window by the red (G12) and green (3-D) dots of Fig. 8. The dashed lines indicate the validation MCC for threshold-based predictions trained on the same, short time-series containing only five events.

The short training data has only a minor effect for predictions of the G12 model with short search windows: the validation MCC of SVM and threshold-based predictions remains high. This is in line with the findings of Gwirtz *et al.* (2021), which suggest that a useful threshold can be determined for G12 from a training data containing only five events. For larger search windows the validation skill scores of SVMs trained on G12 steadily drop off. Larger search windows, however, should have the potential to add information and, therefore increase skill, if there indeed is information contained in the lead-up to a reversal (if there is none, the skill should remain constant with the search window). Thus, an explanation for why the skill score decreases with search window size is that SVMs

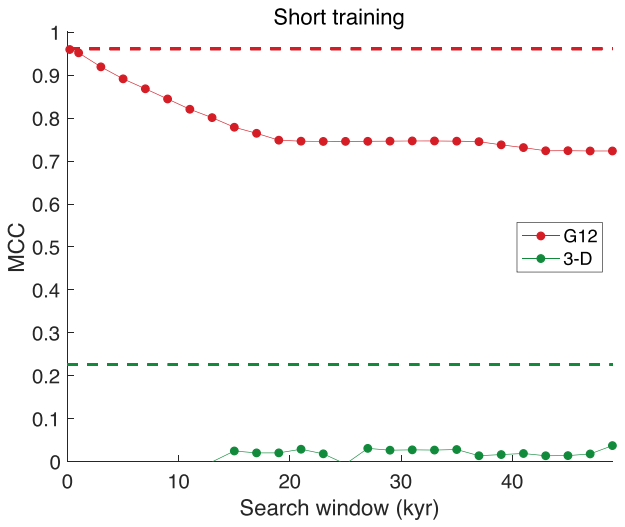

Figure 8. Validation MCC as a function of the search window for SVMs trained on small data sets of the G12 (red dots) and 3-D (green dots) models.

with *limited* training data tend to overfit the training data. Due to overfitting limited training data, the skill scores on *long* validation data drop significantly. Indeed, the long validation data reveals this

overfitting in the form of a small validation MCC. This implies that, for reliable 'learning' via SVMs, one needs a large training data set *and* a large validation data set which, taken together, means one needs 'a lot' of data.

Perhaps unsurprisingly, in the less predictable 3-D model, short training data has devastating effects independently of the search window. All SVMs trained on the short time-series for search windows less than 11 kyr, either always predict an event to occur or, never predict an event to occur (and thus the MCC is undefined). For short search windows, the MCC scores even drop below zero, indicating that one is better served by predicting the *opposite* of what the SVM predicts (on validation data). Similarly, for larger search windows, validation MCC scores for the 3-D model are notably lower than when training with large data. As in the case of G12, the poor validation scores (on long validation data), indicate that the SVMs overfit the short training data. In conclusion, our numerical experiments with the G12 and 3-D models indicate that SVMs overfit short training data and, therefore, lead to unreliable results.

Given the limited palaeomagnetic record (PADM2M and Sint-2000 contain only six low-dipole events) we now test an alternative approach to training and validating SVMs when there is not 'a lot' of data, that is the overall time-series available is short and contains only a few events. Instead of dividing data into two distinct training and validation pieces, we apply stratified K-fold cross validation (see, e.g. Japkowicz & Shah 2011). The procedure is as follows. For a given search window and prediction horizon, all of the segments of a time-series to be used for either training or validation are collected and classified as positives or negatives (see Section 3). These are then randomly sorted into five subsets of equal size and with the same ratio of positives to negatives. One subset is set aside for validation while the remaining data is used for training. The resulting validation MCC is recorded and the training process is repeated four more times, each time using a different subset for validation. The results of applying this process to time-series of G12 and 3-D which contain only six events (just as PADM2M and Sint-2000) are shown in Fig. 9. For each search window, the red and green dots show the average validation MCC scores for G12 and 3-D, respectively. The grey regions span the maximum and minimum MCC values with grey dots representing the maximal/minimal validation scores for each search window. Unfilled diamonds correspond to the maximum MCC achieved during training an SVM on the full time-series (no validation). The dashed lines report the average MCC score of threshold-based predictions (following the same training and validation procedure) for G12 (red) and 3-D (green).

For both models (G12 and 3-D), the average threshold scores and small search window SVM scores are lower than those using long training data. The average score for G12 is generally larger for longer search windows, though it dips down (around a search window size of 45 kyr) and the range of scores continues to span lower values. The average scores of 3-D remain very low for search windows of less than around 35 kyr before increasing to values between 0.2 and 0.3, or around the scores found with large training data (Fig. 6). These results highlight an important lesson for considering machine learning with the limited palaeomagnetic record. Specifically, if only the cross-validation results were available, one might erroneously conclude that, for example, with the 3-D model, the increase in MCC for large search windows indicates that the linear SVMs can find precursors to events. We know however, from the earlier results (Fig. 6), that this is not the case. The issue is that the K-fold cross validation, which relies on short training *and* validation data, cannot reliably reveal an overfitting due to the shortness

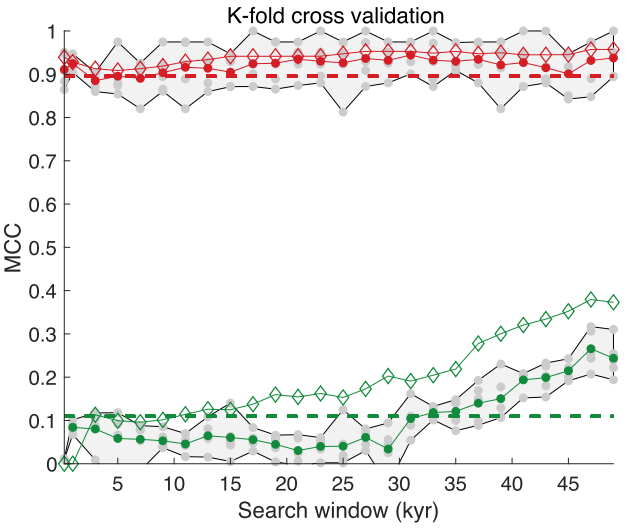

Figure 9. Average MCC validation scores resulting from a stratified fivefold cross validation with a short time-series (six low-dipole events) for G12 (red dots) and 3-D (green dots). The grey regions cover the minimum and maximum validation MCCs with grey dots indicating individual scores. Dashed lines show the average MCC of threshold-based predictions subject to the same stratified fivefold cross validation. The training MCC achieved by fitting the full data is shown as unfilled diamonds.

of the validation data. All investigations of palaeomagnetic reconstructions using machine learning in the above outlined framework, must be done with these limitations in mind. This means in particular, that an increase in (validation) MCC with search window cannot be taken as conclusive evidence that a larger search window indeed leads to better predictions. Rather, the increased (validation) skill may simply be due to the fact that we do not have enough data, for training *and* validation, to reveal overfitting to (short) training data.

4.2 SVMs with palaeomagnetic reconstructions

We now consider the palaeomagnetic reconstructions PADM2M and Sint-2000. To begin, we qualitatively examine the lead-up to events in Fig. 10. Similar to Fig. 7, Fig. 10 shows segments of each times series during the 50 kyr preceding events. For both reconstructions, there is an overall trend of decreasing intensity. This trend is not as consistent as that of the G12 model but is more definitive than anything of a similar timescale in the 3-D model (Fig. 7). Otherwise, there is no obvious (to the human eye) pattern in the behaviour of either time-series prior to events. Moreover, differences between PADM2M and Sint-2000, which could lead to differing ML results, are evident.

To search for precursors using ML, we follow the method outlined in the previous section and apply a stratified fivefold cross validation strategy to the palaeomagnetic reconstructions (because the data we have are too limited to allow for other approaches). Fig. 11 summarizes the results of this process in the same fashion as Fig. 9. For each search window, the dots show the validation MCC scores (grey) and their average (colour). The grey region spans the maximum and minimum validation scores and the unfilled diamonds show the maximum MCC training from an SVM on the full time-series and using no independent validation. The same cross validation strategy is used to determine an average MCC score for threshold-based predictions which is indicated by a dashed line.

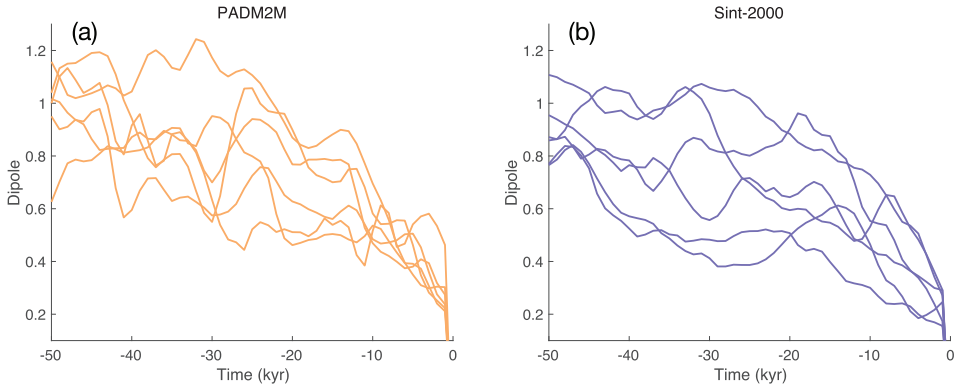

Figure 10. Time-series of 50 kyrs of axial dipole intensity preceding the beginning of events for (a) PADM2M (Ziegler *et al.* 2011) and (b) Sint-2000 (Valet *et al.* 2005).

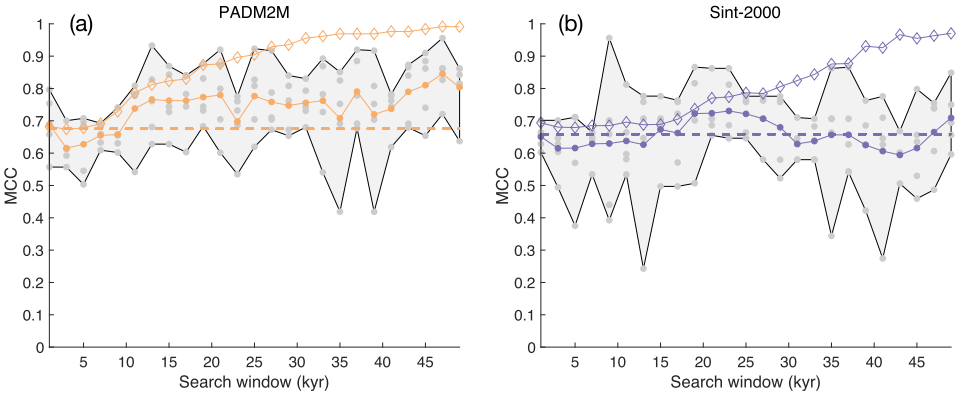

Figure 11. Dots showing individual MCC validation scores (grey) and their average (colour) as a function of search window resulting from stratified fivefold cross validation with (a) PADM2M and (b) Sint-2000. The grey regions cover the minimum and maximum validation scores. Dashed lines show the average MCC of threshold-based predictions subject to the same stratified fivefold cross validation. The training MCC achieved by fitting the full data is shown as unfilled diamonds.

With Fig. 11 in mind one may ask: Are the SVMs detecting precursors of low-dipole events in the observational record of the axial dipole? Unfortunately, as explained in Section 4.1.2 the limited data restricts the level of certainty with which one can draw conclusions. Put simply, we do not have sufficient data to reliably train *and* validate an SVM.

Additionally, as discussed in Section 2.1, there is uncertainty in the palaeomagnetic reconstructions themselves, as evidenced by the differences between PADM2M and Sint-2000. These differences may also lead to discrepancies in the SVM results. Most notably, the average MCC validation scores for PADM2M are consistently better with windows greater than 10 kyrs (similar to the G12 results of Fig. 6) which could suggest the existence of precursors to low-dipole events. However, this is not the case with Sint-2000 (Fig. 11) where, on average, the validation skill score does not change with the search window (similar to the 3-D results of Fig. 6). In our discussion below, we dig deeper into these issues, taking into account the results we obtained with models (where we could more directly study the impact of very limited training/validation data), and with other ML methods (non-linear SVMs and LSTMs).

5 DISCUSSION

We discuss the geophysical significance of our results, in particular with respect to the difficulties that the limited palaeomagnetic data impose on training and validating SVMs. We also investigate the robustness of the results obtained with SVMs by comparing to other ML techniques (non-linear SVMs and LSTMs).

5.1 SVMs and numerical models

The skill of SVMs differs significantly between the G12 and 3-D models. When trained on large data series, SVM predictions of G12 are of high quality (MCC near 1) while predictions of the 3-D model are poor (MCC of around 0.2). This difference in predictability is in line with results of threshold-based predictions (Gwirtz *et al.* 2021). Moreover, the SVMs can pick up on precursors of low-dipole events in the G12 model (skill increases with the search window). When trained on a short data set, we note that the skill of SVMs deteriorates with the search window. This implies that SVMs may be of limited use to search for precursors of low-dipole events when the data are limited, largely because SVMs tend to overfit. Note that our tweak to the training of SVMs addresses imbalances in the data, but cannot address the difficulties arising from the training data being limited.

What is perhaps most notable, however, is not the particular MCC values but the shape of the SVM curves in Fig. 6. When training on long simulations, SVM predictions of the G12 model improve with the search window length, indicating that the SVM can indeed discover (dynamic) precursors of low-dipole events. This is in agreement with Morzfeld *et al.* (2017), where it was found that

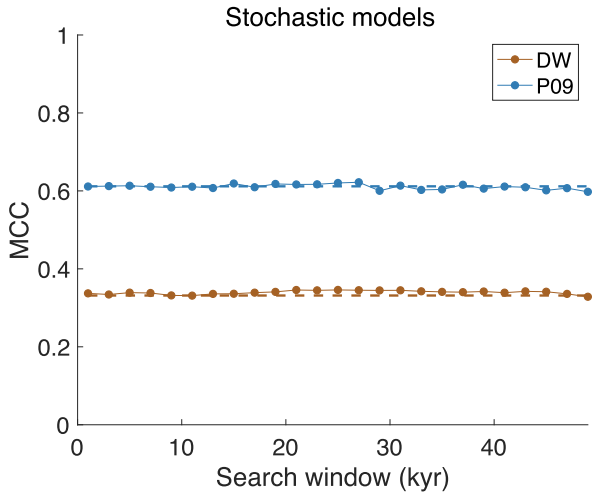

Figure 12. SVM results with the DW (Morzfeld & Buffett 2019) and P09 (Pétrélis *et al.* 2009) models. Dots show validation MCC as a function of the search window length for SVMs trained on a large data set from DW (brown) and P09 (blue). Horizontal dashed lines indicate the MCC of threshold-based predictions using the same training and validation data from DW (brown) and P09 (blue).

predictions with the G12 model were improved when assimilating data over a window of time, that is, when taking dynamics leading up to a reversal into account. The curve of validation MCC scores of SVMs applied to the 3-D model is flat (Fig. 6), indicating that the linear SVMs do not find precursors of low-dipole events.

To understand these findings, we remind the reader that the G12 model is defined by a *system* of three ordinary differential equations, one of which represents the axial dipole (see Section 2.1). Therefore, when SVMs identify precursors of low-dipole events in a time-series of G12 axial dipole intensity, they are finding indicators that the two unobserved components of the system are in a state favourable for causing a low-dipole event. If we consider low-dimensional models which are defined by a single quantity, precursors of low-dipole events may not exist by construction. For example consider stochastic differential equation (SDE) models of the form

$$\mathrm{d}x = f(x)\mathrm{d}t + \sqrt{2q}\,\mathrm{d}W, \tag{4}$$

where $f(x)$ is a prescribed function of x (the drift), q is a constant, W is Brownian motion, and the axial dipole intensity is either x or a simple, deterministic function of x. Because the increments of Brownian motion are uncorrelated in time, the only information useful to predicting future behaviour is the present value of $f(x)$. The SDE models of Morzfeld & Buffett (2019) and Pétrélis *et al.* (2009) take the form of eq. (4) and we now consider these in more detail (the model parameters are as in Gwirtz *et al.* 2021). For short, we refer to the models as DW (short for 'double well', Morzfeld & Buffett 2019) and P09 (Pétrélis *et al.* 2009). In Fig. 12, we show the result of SVMs applied to long training and validation data (180 events each) of the DW model (brown) and P09 model (blue). As in Section 4, the dots represent validation MCC scores from linear SVMs while the dashed lines indicate the validation score of the threshold-based predictions. For both models, the SVM scores do not improve with longer search windows. For DW and P09, this is to be expected because both models are constructed such that, except possibly *during* the occurrence of, or recovery from a low-dipole event, the axial dipole intensity indicates the value of $f(x)$ and therefore, no additional information is obtained by looking into the past. Indeed, it appears as if the axial dipole time-series of the 3-D model essentially behaves like the solution of a scalar SDE model (in terms of predictability via SVMs).

5.2 SVMs and palaeomagnetic reconstructions

The limited length of the palaeomagnetic reconstructions of PADM2M and Sint-2000 makes it difficult to interpret the results obtained by applying SVMs to search for precursors of magnetic reversals. The linear SVM experiments of Section 4.2 do not show convincing evidence of precursors of reversals in the palaeomagnetic record for the following two reasons:

(i) Average skill of SVMs is different for PADM2M and Sint-2000.
(ii) The limited data cause large uncertainties in the skill scores.

Differences in average skill may be due to the fact that the two reconstructions rely on quite different numbers of 'raw' data and process these data selections in different ways. The large uncertainties in Fig. 11, however, are critical and do not allow for the conclusion that skill improves with the search window, that is, the SVMs are ultimately not capable of detecting precursors of reversals. Additionally, we note that we must be carefu because, even if SVMs were capable of detecting precursors, these may not necessarily reflect a property of the Earth's axial dipole, but, instead, could be the result of the smoothing of the time-series due to the sedimentary processes through which the past field is recorded, limited age control of individual records, or the mathematical methods used to reconstruct that record.

If there are indeed no precursors of low-dipole events in the palaeomagnetic reconstructions, then the use of SDE models such as DW and P09 with the same property may be justified as a basic stochastic field descriptor. This is in line with the results of Gwirtz *et al.* (2021) where it was found that P09 had the most 'Earth-like' properties with respect to threshold-based predictions.

Overall, our results suggest that threshold-based predictions may define a limit for how well one can anticipate Earth's low-dipole events, from only a limited observational record. The more sophisticated SVMs, even tweaked to work with imbalanced data, tend to overfit when training data is limited, which means that SVM-based predictions may be flawed and are ultimately not able to reliably discover precursors of low-dipole events of the Earth within the currently available PADM records.

5.3 Robustness of results

We test the robustness of our SVM methodology by repeating a limited set of numerical experiments using two additional ML techniques. We first consider a non-linear SVM. As indicated in Section 2.3, the idea is to apply a transformation to the (training) data that may make it linearly separable by an SVM. We use radial basis functions for the transformation, which is a common choice (RBFs, see, e.g. Cristianini & Shawe-Taylor 2000). This is accomplished by using the kernel function option `rbf` in Matlab's `fitcsvm` function. The top row of Fig. 13 shows the result of applying a non-linear SVM to the palaeomagnetic reconstructions using the same stratified, fivefold cross validation of Section 4.2. The dots show the validation MCC scores (grey) and their average (colour) with the grey cloud spanning the minimum and maximum validation scores. The unfilled diamonds show the training MCC of the non-linear

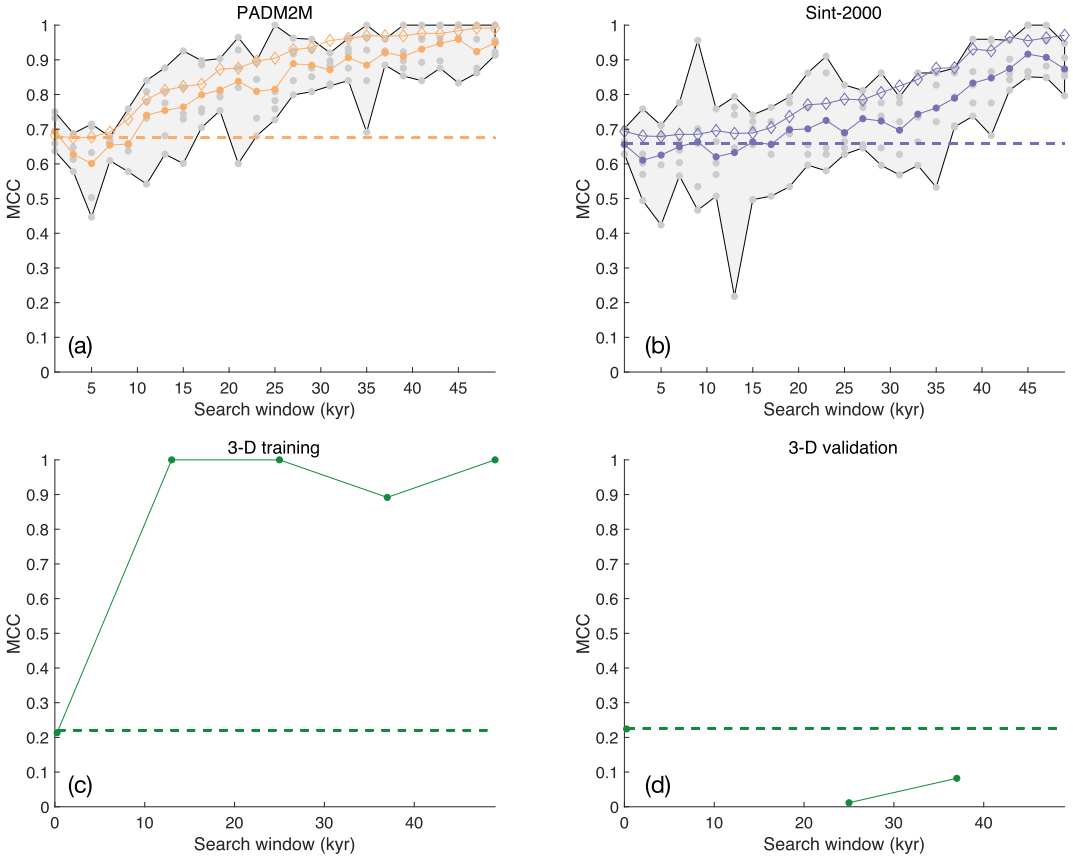

Figure 13. Top row: dots showing individual MCC validation scores (grey) and their average (colour) as a function of search window resulting from stratified fivefold cross validation with (a) PADM2M and (b) Sint-2000 using non-linear SVMs. The grey regions cover the minimum and maximum validation scores. The dashed lines show the average MCC of threshold-based predictions subject to the same stratified fivefold cross validation. The training MCC achieved by fitting the full data is shown as unfilled diamonds. Bottom row: (c) training and (d) validation MCC using large training and validation data sets of the 3-D model with non-linear SVMs (dots) and the threshold strategy (dashed lines).

SVM (applied to the full data sets). Dashed lines show the average validation MCC of threshold-based predictions, using the same stratified fivefold cross validation. We now see that skill increases with search window, which suggests that the non-linear SVMs may have discovered precursors of low-dipole events.

Using the numerical models, however, we can show that the RBFs allow for a level of overfitting which makes the results unreliable. Specifically, the bottom row of Fig. 13 shows training and validation scores using non-linear SVMs with a long simulation of the 3-D model with long training *and* long validation data (the training/validation data as in Fig. 6, so that the validation can reveal an overfitting, see Section 4.1.2). We see that the SVM obtains training MCCs near 1, but this skill does not generalize to long validation data. Indeed, when inspecting skill on validation data [panel (d) of Fig. 13], it becomes clear that the non-linear SVM has overfit. For several search windows, the non-linear SVM fails to predict any of the low-dipole events in the validation data (so that MCC is undefined and thus, not reported), and even when defined, MCC on validation data is very low (below threshold-based predictions or linear SVMs). This is in contrast to the linear SVMs, which exhibit roughly the same training and validation scores when trained and validated with long data (the training MCC, not shown in Fig. 6, is comparable to the validation MCC).

Finally, we consider the use of long short-term memory networks (LSTMs, see Hochreiter & Schmidhuber 1997). We perform the same set of experiments as with the non-linear SVMs (above), that is, stratified fivefold cross validation with the palaeomagnetic reconstructions and large training and validation data sets with the numerical models. The results are shown in Fig. 14 where we again find no clear evidence of low-dipole event precursors in the palaeomagnetic reconstructions (top row). Indeed, with each reconstruction, for multiple search windows there are validation runs during which no low-dipole events are predicted by the LSTM, resulting in undefined MCC scores, and thus no averages are plotted. Most notably, we see in the bottom row of Fig. 14 that results with the 3-D model are similar to those with linear SVMs in that they are unaffected by the search window. This further supports the conclusion that the axial dipole intensity time-series of the 3-D model contains no precursors of low-dipole events.

5.4 Beyond ML and threshold-based predictions

To gain further insights into how to interpret the results from the ML techniques, we study the stochastic properties of the axial dipole time-series of the various simulations and palaeomagnetic reconstructions *without* SVMs, other ML techniques, or threshold-based predictions. Specifically, for a long simulation of G12, DW, P09, the full simulation of the 3-D model and the entirety of PADM2M and Sint-2000, we compute the time-series of *change* in axial dipole

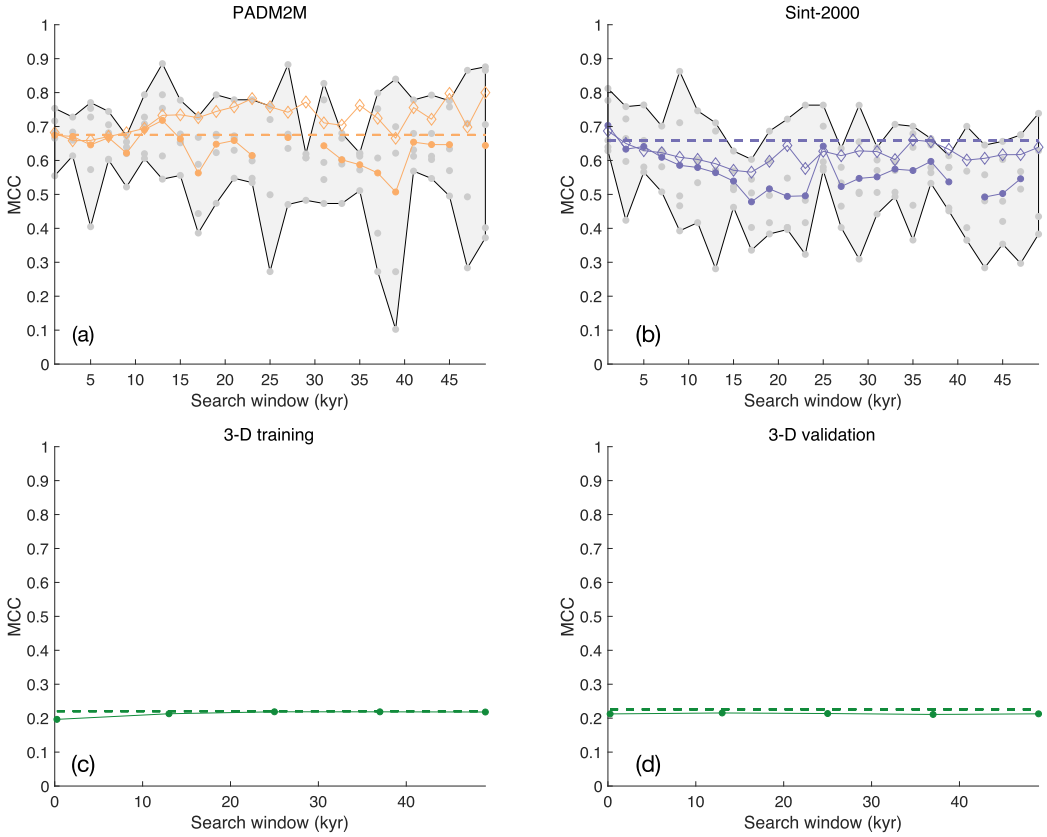

Figure 14. Top row: dots showing individual MCC validation scores (grey) and their average (colour) as a function of search window resulting from stratified fivefold cross validation with (a) PADM2M and (b) Sint-2000 using LSTMs. Search windows for which the LSMTs predicted no events result in an undefined MCC. The grey regions cover the minimum and maximum validation scores. The dashed lines show the average MCC of threshold-based predictions subject to the same stratified fivefold cross validation. The training MCC achieved by fitting the full data is shown as unfilled diamonds. Bottom row: (c) training and (d) validation MCC using long simulations of the 3-D model with LSTMs (dots) and the threshold strategy (dashed lines).

intensity, defined as first differences. If the axial dipole intensity at time i is d_i, the time-series of change in axial dipole intensity is $\Delta d_i = d_{i+1} - d_i$. The idea is to use this time-series to determine if, at any given time, correlations exist between the change that is about to occur in axial dipole intensity, and previous changes.

The autocorrelation function of the time-series of change in axial dipole intensity is shown for the models and palaeomagnetic reconstructions in Fig. 15. Here, we resampled the model outputs so that the temporal resolution is approximately equal to that of the palaeomagnetic reconstructions (1 kyr). We see that G12 exhibits strong correlations, with the change in axial dipole intensity over the coming 1 kyr being strongly positively correlated with the change over the previous 1 kyr (correlation coefficient greater than 0.8). This perhaps highlights one of the reasons why SVM predictions proved most useful with the G12 model. For lags of less than 10 kyr, the pattern of autocorrelations for G12 most closely resemble those of the palaeomagnetic reconstructions, which also indicate correlation between future and past changes, but perhaps to a weaker degree, reflecting the inevitable decline in quality of the palaeofield records on these short timescales. Interestingly, but perhaps not surprisingly given the results of our SVM experiments, changes in the axial dipole intensity of the 3-D model are largely uncorrelated over timescales of 1 kyr. Indeed, with its absence of any large-magnitude correlations, the 3-D model appears most similar to the SDE models of DW and P09 for which we know future changes are independent of the past.

In summary, the above considerations support the conclusions we draw from using SVMs (and threshold-based predictions) to study the models and the data:

(i) The axial dipole of the 3-D model does not contain precursors of low-dipole events. This is supported by the fact that autocorrelation of the time-series of change in axial dipole intensity of the 3-D model is similar to the autocorrelation of a stochastic model, for which we know that no precursors exist.

(ii) The axial dipole of G12 contains precursors—and the SVMs are able to detect these, given sufficient training data. This is supported by the autocorrelation of the time-series of change in axial dipole intensity of G12.

(iii) The palaeomagnetic reconstructions exhibit characteristics lying between the G12 and the stochastic models: The skill of ML methods applied to the palaeomagnetic record fall (quantitatively) between the corresponding scores of the G12 and the stochastic models. Similarly, the 'shape' of the auto correlation functions of the palaeomagnetic records is somewhat 'in between' the shapes of the stochastic models (nearly no correlation) and the G12 model (strong correlation).

We note that one can also inspect power spectral densities (PSD) of the axial dipole intensities and note differences between the G12 model and the other models (3-D, P09 and DW), and the palaeomagnetic reconstructions. This suggests that one can possibly

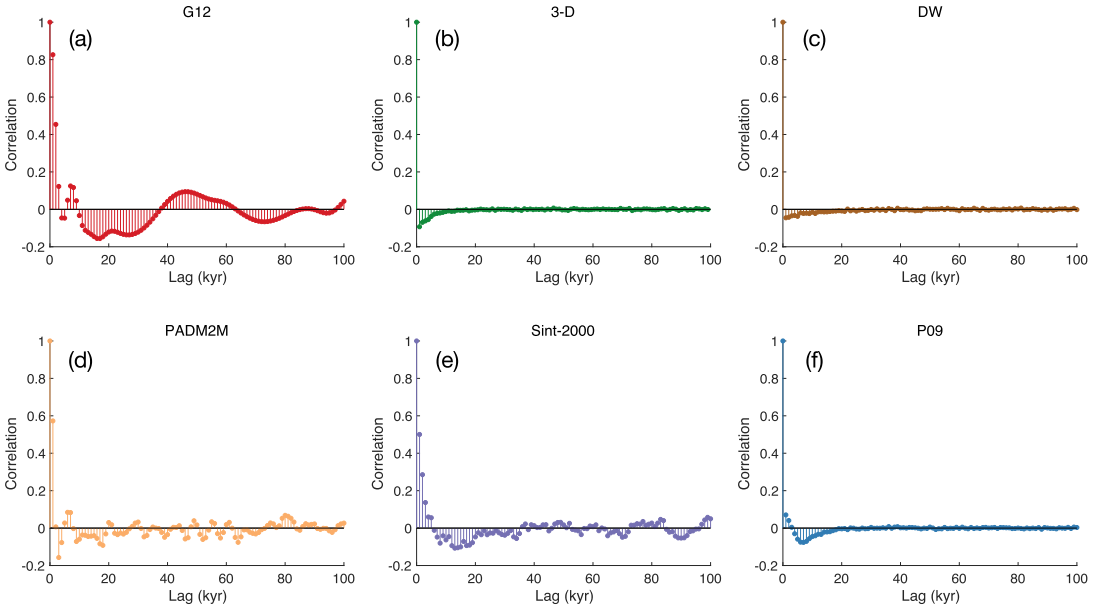

Figure 15. Autocorrelation functions of time-series of change in axial dipole intensity for (a) G12, (b) 3-D, (c) DW, (d) PADM2M, (e) Sint-2000 and (f) P09.

make connections between PSDs and predictability, but we do not pursue these ideas any further here.

6 CONCLUSIONS

We applied machine learning methods to search for dynamic precursors of reversals and major excursions (collectively, low-dipole events) of Earth's axial dipole field. We benchmarked the ML techniques and implied predictions against a simpler threshold-based strategy which does not take dynamics into account. To make this possible, we equipped some standard ML tools with a tweak to more robustly handle imbalanced data—data where one event occurs much more frequently than another (no low-dipole event versus low-dipole event). Studying low-dipole events of Earth is difficult because data are limited. We address this issue by invoking a hierarchy of models for Earth's axial dipole. By training and validating ML methods on model output, which is not subject to any limitation on the amount of data, we can study the effects of limited data on ML techniques, which helps with interpreting the results obtained when applying ML to the limited set of data we have. Our main findings are as follows.

(i) ML is robustly capable of identifying precursors of low-dipole events of the G12 *model*. Taking the dynamics preceding a low-dipole event into account, the ML could indeed perform more accurately than a simpler threshold-based strategy that does not take dynamics into account. This increase in predictive capability due to ML, however, is conditioned on large training data sets and is somewhat minor, because even simple threshold-based strategies already lead to accurate predictions. Moreover, the simpler threshold-based strategy is more robust in view of limited training data and, for these reasons, perhaps preferable to sophisticated ML methods.

(ii) The axial dipole time-series of a 3-D numerical dynamo model does *not* contain dynamic precursors of low-dipole events. We arrived at this conclusion by applying several ML techniques (so that we do not overlook precursors by choosing inappropriate algorithms) and by detailed comparisons to scalar stochastic differential equation (SDE) models which, by construction, do not contain dynamic precursors. If the 3-D simulation accurately represents the characteristics of Earth's axial dipole evolution, then this implies that one need not search for dynamic precursors of low-dipole events in Earth's axial dipole time evolution. As an aside, our study suggests that the axial dipole of this 3-D simulation can be modelled by a scalar SDE (but we cannot rule out that other simulations may exhibit different characteristics).

(iii) We did not find convincing evidence of dynamic precursors of low-dipole events in two palaeomagnetic reconstructions. This does not rule out that such precursors may exist, but collectively our study makes a strong case that the current reconstructions of axial dipole dynamics leading up to a reversal do not contain significant information that renders viable ML predictions of an upcoming reversal or major excursion.

We emphasize that our study does not allow us to rule out the existence of dynamic precursors of Earth's low-dipole events, or that other numerical techniques may be able to discover these. More sophisticated ML methods, however, typically require a greater amount of training data to constrain a large number of parameters within the ML algorithm. It is therefore unlikely that more sophisticated ML techniques can robustly identify precursors of low-dipole events in PADM2M or Sint-2000. However, currently available shorter high resolution reconstructions of axial dipole dynamics and their power spectra suggest that the geomagnetic field may be more similar to the G12 models in its spectral behaviour than to the 3-D model, leaving room for hope that future palaeomagnetic work may ultimately enable successful identification of precursory behaviour. Another way forward is to focus on realistic numerical models and supply ML methods with features of the magnetic field beyond the axial dipole. This may increase our understanding of the dynamics leading up to a reversal or major excursion. Direct application of these ideas to Earth's magnetic field, however, will remain difficult because we have only limited data of the past (or even present) state of the geodynamo, beyond its large-scale features. Moreover, geodynamo models are routinely run with parameters far from those speculated to be realistic for Earth's dynamo (due to computational constraints), so even 'realistic' model outputs should

be interpreted carefully. Nonetheless, our experiments and explanations indicate that ML techniques, despite their limitations, may be useful for improving our understanding of reversals of Earth's magnetic field and we hope that our work sparks interest in these ideas.

ACKNOWLEDGMENTS

KG was supported by an appointment to the NASA Postdoctoral Program at Goddard Space Flight Center, administered by Oak Ridge Associated Universities under contract with NASA. TD was supported by a Summer Undergraduate Research Fellowship (SURF), awarded by Scripps Institution of Oceanography, University of California, San Diego. MM and KG were supported by the US Office of Naval Research (ONR) grant N00014-21-1-2309. CC was supported by CC was supported by NSF grant EAR 1953778. AF was supported by the French Agence Nationale de la Recherche under grant ANR-19-CE31-0019 (revEarth).

All authors contributed to the ideas presented in this paper with KG taking the lead and writing the first draft. KG, TD and MM wrote the code.

DATA AVAILABILITY

Code for implementing the machine learning strategies described is available on github (https://github.com/kjg136/MLdipolePredictions). The code and numerical results used to generate the figures have been archived at (https://zenodo.org/record/6568036#.Yof1bi-B30o).

REFERENCES

Batuwita, R. & Palade, V., 2013. *Class Imbalance Learning Methods for Support Vector Machines,* Chapter 5, pp. 83–99, John Wiley & Sons, Ltd..

Ben-Hur, A., Ong, C.S., Sonnenburg, S., Schölkopf, B. & Rätsch, G., 2008. Support vector machines and kernels for computational biology, *PLOS Comput. Biol.,* **4**(10), 1–10.

Brown, M., Korte, M., Holme, R., Wardinski, I. & Gunnarson, S., 2018. Earth's magnetic field is probably not reversing, *Proc. Natl. Acad. Sci.,* **115**(20), 5111–5116.

Cande, S. & Kent, D., 1995. Revised calibration of the geomagnetic polarity timescale for the late cretaceous and Cenozoic, *J. geophys. Res.,* **100,** 6093–6095.

Chicco, D. & Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genom.,* **21**(1), 6.

Constable, C. & Korte, M., 2006. Is Earth's magnetic field reversing?, *Earth planet. Sci. Lett.,* **246,** 1–16.

Cortes, C. & Vapnik, V., 1995. Support-vector networks, *Mach. Learn.,* **20**(3), 273–297.

Cristianini, N. & Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods,* Cambridge Univ. Press.

Gissinger, C., 2012. A new deterministic model for chaotic reversals, *Eur. Phys. J. B.,* **85,** 137, doi:10.1140/epjb/e2012-20799-5.

Goodfellow, I., Bengio, Y. & Courville, A., 2016. *Deep Learning,* MIT Press, http://www.deeplearningbook.org.

Graves, A., 2012. *Supervised Sequence Labelling with Recurrent Neural Networks, Studies in Computational Intelligence,* Springer Berlin Heidelberg.

Gwirtz, K., Morzfeld, M., Fournier, A. & Hulot, G., 2021. Can one use Earth's magnetic axial dipole field intensity to predict reversals?, *Geophys. J. Int.,* **225**(1), 277–297.

Hochreiter, S. & Schmidhuber, J., 1997. Long short-term memory, *Neural Comput.,* **9**(8), 1735–1780.

Hulot, G., Lhuillier, F. & Aubert, J., 2010. Earth's dynamo limit of predictability, *Geophys. Res. Lett.,* **37**(6), doi:10.1029/2009GL041869.

Japkowicz, N. & Shah, M., 2011. *Evaluating Learning Algorithms: A Classification Perspective,* Cambridge Univ. Press.

Kim, K., 2003. Financial time series forecasting using support vector machines, *Neurocomputing,* **55**(1), 307–319.

Kok, Z.H., Mohamed Shariff, A.R., Alfatni, M. S.M. & Khairunniza-Bejo, S., 2021. Support vector machine in precision agriculture: a review, *Comp. Electr. Agric.,* **191,** doi:10.1016/j.compag.2021.106546.

Laj, C. & Kissel, C., 2015. An impending geomagnetic transition? Hints from the past, *Front. Earth Sci.,* **3,** 61, doi:10.3389/feart.2015.00061.

Lhuillier, F., Aubert, J. & Hulot, G., 2011a. Earth's dynamo limit of predictability controlled by magnetic dissipation, *Geophys. J. Int.,* **186,** 492–508.

Lhuillier, F., Fournier, A., Hulot, G. & Aubert, J., 2011b. The geomagnetic secular-variation timescale in observations and numerical dynamo models, *Geophys. Res. Lett.,* **38**(9), doi:10.1029/2011GL047356.

Lowrie, W. & Kent, D., 2004. Geomagnetic polarity time scale and reversal frequency regimes, *Geophys. Monogr. Ser.,* **145,** 117–129.

Ma, Y. & Guo, G., 2014. *Support Vector Machines Applications,* Springer International Publishing.

Morzfeld, M. & Buffett, B.A., 2019. A comprehensive model for the kyr and Myr timescales of Earth's axial magnetic dipole field, *Nonlin. Proc. Geophys.,* **26**(3), 123–142.

Morzfeld, M., Fournier, A. & Hulot, G., 2017. Coarse predictions of dipole reversals by low-dimensional modeling and data assimilation, *Phys. Earth planet. Inter.,* **262,** 8–27.

Murty, M. & Raghava, R., 2016. *Support Vector Machines and Perceptrons: Learning, Optimization, Classification, and Application to Social Networks, SpringerBriefs in Computer Science,* Springer International Publishing.

Ogg, J., 2012. Geomagnetic polarity time scale, in *The Geologic Timescale,* Chapter 5, pp. 85–113, eds Gradstein, F., Ogg, J., Schmitz, M. & Ogg, G., Elsevier Science.

Olson, P., Driscoll, P. & Amit, H., 2009. Dipole collapse and reversal precursors in a numerical dynamo, *Phys. Earth planet. Inter.,* **173**(1), 121–140.

Pétrélis, F., Fauve, S., Dormy, E. & Valet, J.-P., 2009. Simple mechanism for reversals of Earth's magnetic field, *Phys. Rev. Lett.,* **102,** 144503.

Platt, J., 1998. Sequential minimal optimization: a fast algorithm for training support vector machines, Tech. Rep. MSR-TR-98-14, Microsoft.

Valet, J.-P. & Fournier, A., 2016. Deciphering records of geomagnetic reversals, *Rev. Geophys.,* **54**(2), 410–446.

Valet, J.-P., Meynadier, L. & Guyodo, Y., 2005. Geomagnetic field strength and reversal rate over the past 2 million years, *Nature,* **435,** 802–805.

Ziegler, L.B., Constable, C.G., Johnson, C.L. & Tauxe, L., 2011. PADM2M: a penalized maximum likelihood model of the 0-2 Ma paleomagnetic axial dipole model, *Geophys. J. Int.,* **184**(3), 1069–1089.

APPENDIX: BACKGROUND ON SUPPORT VECTOR MACHINES

We provide additional detail on how (linear) support vector machines (SVM) solve a binary classification problem and how we tweak Matlab code to be able to deal with imbalanced data.

Consider a collection of data of the form $(\mathbf{x}_i, y_i)$, for $i = 1, \ldots k$ where $\mathbf{x}_i \in \mathbf{R}^n$ and $y_i \in \{1, -1\}$ indicates which of two classes the vector $\mathbf{x}_i$ belongs to. The $+1$ here could be the class of 'cats' (or time-series chunks followed by a low-dipole event), and the -1 could identify 'dogs' (or time-series chunks *not* followed by a low-dipole event). The hyperplane that separates the two classes can be parametrized by the set of all points $\mathbf{x} \in \mathbf{R}^n$ such that $\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$ for some $\mathbf{w} \in \mathbf{R}^n$, and scalar b.

If the data are linearly separable, a SVM finds the separating hyperplane by solving the optimization problem

$$\min_{\mathbf{w},b} \quad \left(\frac{1}{2}||\mathbf{w}||_2^2\right), \tag{A1}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1, \text{ for all } i = 1, \ldots k, \tag{A2}$$

(Cortes & Vapnik 1995). The above is to be understood in the following way. The unknowns are the n elements of the vector $\mathbf{w}$ and the scalar b. These $n + 1$ parameters define the separating hyperplane and are determined by minimizing the 2-norm of $\mathbf{w}$, subject to (s.t.) the k (number of training data) constrains in (A2); in (A1), the vertical bars denote the two norm of a vector, for example if $\mathbf{w}$ is an n-dimensional vector with elements w_i, $i = 1, \ldots, n$, then $||\mathbf{w}||_2 = \sqrt{\sum_{i=1}^{n} w_i^2}$.

If the data are *not* linearly separable, as is the case in almost all problems, including our study, the SVM determines a hyperplane by solving the optimization problem

$$\min_{\mathbf{w},b} \left(\frac{1}{2}||\mathbf{w}||_2^2 + \frac{1}{2}\sum_{i=1}^{k}[C_{\mathrm{FN}}(1 + y_i) + C_{\mathrm{FP}}(1 - y_i)]\xi_i\right) \tag{A3}$$

with

$$\xi_i = \max(0, 1 - y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b)) \tag{A4}$$

for all $i = 1,\ldots k$ where C_{FN} and C_{FP} are constants and where ξ_i are 'slack variables'. (Batuwita & Palade 2013). The slack variables ξ_i are a measure of the extent of misclassifications by a candidate hyperplane and the constants C_{FN} and C_{FP} determine the size of penalty assigned to false negatives (incorrectly assigning the class of $y = -1$) and false positives (incorrectly assigning the class of $y = 1$), respectively.

We use Matlab's `fitcsvm` function, and we note that `fitcsvm` performs the required optimization by considering the dual Lagrangian form of (A3). Numerically, the optimization is implemented via the sequential minimal optimization algorithm (Platt 1998).

Our tweak of using SVMs on imbalanced data is implemented as follows. Matlab's `fitcsvm` function allows for (limited) modification through user input, which we use to vary the ratio $r = C_{\mathrm{FP}}/C_{\mathrm{FN}}$ (the FP/FN penalty ratio) within the code. For a fixed value of r, the constants are $C_{\mathrm{FN}} = (\mathrm{P} + \mathrm{N})/(\mathrm{P} + r\mathrm{N})$ and $C_{\mathrm{FP}} = r(\mathrm{P} + \mathrm{N})/(\mathrm{P} + r\mathrm{N})$, where P and N are the number of positives and negatives in the training data, respectively. Here, we determine one SVM for each value of r (on a grid), compute the resulting MCC over the training data, and declare the SVM that maximizes MCC as optimal.

Finally, we provide detail of how the training data are composed when searching for precursors of low-dipole events. Suppose that $n = 1$. Then the training data $\mathbf{x}_i$ are scalars and are simply the elements of the time-series of the axial dipole. The labels indicate whether or not each element (time instance) is followed by a low-dipole event within the prediction horizon or not. In this case, the SVM essentially determines a threshold. If $n = 2$, the training data are 2-D vectors $\mathbf{x}_i$ that contain two consecutive axial dipole intensities. The first element of each $\mathbf{x}_i$ is the 'current' axial dipole intensity and the second element is the axial dipole intensity one time step prior to the first element. The labels are as before and indicate whether or not the sequence $\mathbf{x}_i$ is followed by a low-dipole event within the prediction horizon or not. This train of thought generalizes to $n > 2$, for which each $\mathbf{x}_i$ contains n consecutive axial dipole intensities.