UNIVERSITY OF CALIFORNIA SAN DIEGO

Computational tools for estimating and predicting the state of the geodynamo

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Earth Sciences

by

Kyle Gwirtz

Committee in charge:

Professor Matthias Morzfeld, Chair
Professor Catherine Constable
Professor Bruce Cornuelle
Professor Jeffrey Gee
Professor Melvin Leok

2021

The dissertation of Kyle Gwirtz is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

Dissertation Approval Page iii

Table of Contents iv

List of Figures vii

List of Tables ix

Acknowledgements x

Vita xii

Abstract of the Dissertation xiii

Chapter 1 Introduction 1
1.1 The geodynamo 1
1.2 Data assimilation and its application to geomagnetism 3
1.2.1 A brief review of data assimilation 3
1.2.2 Localization and inflation 8
1.2.3 Developing DA with proxy models 10
1.2.4 Geomagnetic DA 11
1.3 Predicting reversals of the Earth's magnetic dipole 14
1.3.1 Reversal prediction as a yes or no question 14
1.3.2 Prediction strategies 15
1.3.3 Reversal predictions as a tool for characterizing dipole behavior 17
1.4 Summary 17

Chapter 2 A testbed for geomagnetic data assimilation 24
2.1 Introduction 25
2.2 A proxy model for geomagnetic DA 27
2.2.1 Background on geodynamo models and geomagnetic DA . . 27
2.2.2 A wishlist of properties for a useful proxy model 29
2.2.3 Mathematical derivation of the proxy model 30
2.2.4 Properties of the proxy 36
2.2.5 Summary of the proxy model 44
2.3 Using the proxy to study geomagnetic ensemble DA 44
2.3.1 Review of data assimilation and the ensemble Kalman filter 45
2.3.2 Keeping the ensemble size manageable: Localization and inflation 48
2.3.3 Localization schemes for geomagnetic DA 50
2.3.4 Observing system simulation experiments (OSSEs) 55
2.4 Results of data assimilation experiments 56

2.4.1 Observation network and metrics of success 56
2.4.2 Results: proxy on the square 59
2.4.3 Results: proxy on the sphere 72
2.5 Summary and conclusions 75

Chapter 3 Can one use Earth's magnetic axial dipole field intensity to predict reversals? 83
3.1 Introduction 84
3.2 Background: model hierarchy and skill scores 87
3.2.1 Numerical modeling of the geomagnetic field 87
3.2.2 The model hierarchy 88
3.2.3 Similarities and differences across the model hierarchy . . . 94
3.2.4 Predictions, skill scores, and ROC curves 97
3.3 Finding thresholds for the prediction of low-dipole events 100
3.3.1 Precise formulation of threshold-based predictions 101
3.3.2 Scaling thresholds and re-scaling time 103
3.3.3 Finding thresholds via maximization of skill scores 107
3.4 Application to a hierarchy of models 107
3.4.1 Skill of threshold-based predictions 108
3.4.2 Robustness of skill to a short training period 116
3.4.3 Impact of data filtering 117
3.4.4 Summary of results from the hierarchy of models 120
3.5 Application to paleomagnetic reconstructions 122
3.5.1 Event durations and decay times 123
3.5.2 Threshold-based predictions and their skills 125
3.6 Concluding comments 131

Chapter 4 Can machine learning find precursors to reversals of Earth's magnetic dipole field? 142
4.1 Introduction 143
4.2 Background: Models, paleomagnetic reconstructions, threshold-based predictions and machine learning 144
4.2.1 Numerical models and paleomagnetic reconstructions 145
4.2.2 Building on previous work in threshold-based predictions . . 146
4.2.3 Machine learning methods 150
4.3 Training SVMs for imbalanced data 151
4.4 SVM prediction results 153
4.4.1 Results with the G12 and 3-D numerical models 154
4.4.2 SVMs with paleomagnetic reconstructions 156
4.5 Discussion 158
4.5.1 Discussion of experiments with SVMs and numerical models 158
4.5.2 Discussion of experiments with SVMs and paleomagnetic reconstructions 160
4.5.3 Sensitivity of results to the choice of machine learning model 161

4.6 Concluding remarks 165
4.6.1 Outlook 166

Chapter 5 Concluding comments 170
5.1 Summary 170
5.2 Outlook 171

LIST OF FIGURES

Figure 1.1: An illustration of an EnKF applied to a system of two variables. 5
Figure 1.2: An illustration of a Gaussian decay of correlations. 9
Figure 1.3: Solutions of the Lorenz 96 and Lorenz 63 sysetms. 11
Figure 1.4: Paleomagnetic reconstruction PADM2M with sign indicating polarity and red horizontal lines indicating a threshold intensity below which one might predict a reversal. 15
Figure 1.5: An SVM classifying data after being trained on a simulation of the magnetic axial dipole. 16

Figure 2.1: Snapshot of a solution of the proxy model on the square and sphere 38
Figure 2.2: PSDs of the proxy model 39
Figure 2.3: Timescales of the proxy model 41
Figure 2.4: Exponential error growth of proxy model 43
Figure 2.5: Climatological correlations of the proxy model 53
Figure 2.6: Forecast errors as a function of ensemble size when using no localization or inflation. 60
Figure 2.7: Forecast errors as a function of ensemble size when using localization and inflation. 61
Figure 2.8: Forecast errors during spin-up. 66
Figure 2.9: Localization and inflation tuning grid. 67
Figure 2.10: Forecast error correlations during a large ensemble run. 71
Figure 2.11: A comparison of forecasts produced with shrinkage and climatological localization schemes. 74

Figure 3.1: Dipole as a function of time for the four models considered in this study. . 95
Figure 3.2: Scaled histograms of the dipole intensities of the four models and two paleomagnetic reconstructions. 96
Figure 3.3: Three examples of ROC curves. 100
Figure 3.4: Excerpt of the dipole of the 3D simulation. 101
Figure 3.5: Illustration of the threshold prediction strategy. 104
Figure 3.6: ROC curves for the four models and three prediction horizons. 109
Figure 3.7: Illustration of threshold-based predictions for the four models. 111
Figure 3.8: MCC skill scores of the four models. 113
Figure 3.9: Average decay time vs. average event duration. 115
Figure 3.10: Training and verification MCC as a function of training data. 117
Figure 3.11: MCC vs. warning threshold for the four models. 118
Figure 3.12: Average decay time plotted as a function of the average event duration for the four models and the paleomagnetic reconstructions. 124
Figure 3.13: ROC curves and MCC as a function of warning threshold for the paleomagnetic reconstructions. 126
Figure 3.14: Verification MCC of four models and two paleomagnetic reconstructions. . 127

Figure 3.15: Illustration of threshold-based predictions for the PADM2M Sint-2000. . . 129

Figure 4.1: Dipole intensity as a function of time for the G12 and 3-D models 146
Figure 4.2: Paleomagnetic reconstructions of the VADM. 147
Figure 4.3: An illustration of the machine learning prediction strategy. 148
Figure 4.4: Training machine learning models to maximize MCC. 153
Figure 4.5: Validation MCC for SVMs trained on long simulations. 154
Figure 4.6: Validation MCC for SVMs trained on short simulations. 156
Figure 4.7: MCC scores using linear SVMs and stratified k-fold cross validation with the paleomagnetic reconstructions. 157
Figure 4.8: MCC for SVMs trained on long simulations of stochastic models. 160
Figure 4.9: Autocorrelation functions of time series of change in dipole intensity for numerical models and paleomagnetic reconstructions. 162
Figure 4.10: MCC scores using non-linear SVMs. 163
Figure 4.11: MCC scores using LSTMs 164

LIST OF TABLES

Table 2.1: Parameter values for the proxy model on a square and sphere. 37
Table 2.2: The ratios of the L^2 norms of the Lorentz force and linear forcing, and the induction and magnetic diffusion terms. 44
Table 2.3: Average forecast errors for the proxy model on the square when using various localization schemes. 62
Table 2.4: Long-range forecast errors resulting from using various localization schemes for the proxy model on the square. 64
Table 2.5: Assimilation cycles required to reduce forecast errors in the magnetic field and velocity field to 1% of the macroscopic error for the proxy model on the square. 67
Table 2.6: Average forecast errors for the proxy model on the sphere. 73
Table 2.7: Assimilation cycles required to reduce forecast errors in the magnetic field and velocity field to 1% of the macroscopic error for the proxy model on the sphere. 74

Table 3.1: Main table summarizing key results obtained throughout the paper. 106
Table 3.2: Maximum MCC of threshold-based predictions for the DW and 3D models with and without smoothing. 120
Table 3.3: Acronyms used in this paper 137

Table 4.1: Prediction horizons used for the models and paleomagnetic reconstructions. 149

ACKNOWLEDGEMENTS

I would like to begin by expressing my deepest gratitude to my advisor Matti Morzfeld. I have been extraordinarily fortunate to have Matti as a mentor over the last few years and will be forever grateful for all of the time and energy he has put into helping me learn and grow. Thank you Matti for all of your guidance and for all of the opportunities you have provided.

Thank you also to the members of my committee. I am sorry that we have not been able to gather together in person during this unusual period. I greatly appreciate that in spite of the circumstances, you have each made time to meet virtually and provide me with meaningful feedback. In particular, I want to thank Cathy Constable, who has been a great source of advice, insight, and encouragement.

I have had the good fortune to work with a number of people who have gone out of their way to help me be successful. Among them, I would especially like to thank Weijia Kuang, Andy Tangborn, Alex Fournier and Gauthier Hulot. Each has been exceptionally generous with their time and has taught me a great deal.

I would like to acknowledge that this work was supported by NASA Headquarters under the NASA Earth and Space Science Fellowship Program - Grant “80NSSC18K1351”.

———————

Chapter 2, in full, is a reprint of material as it appears in Gwirtz, K., Morzfeld, M., Kuang, W. and Tangborn, A., A testbed for geomagnetic data assimilation, *Geophysical Journal International*, **227** (3), 2180-2203, (2021). The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of material as it appears in Gwirtz, K., Morzfeld, M., Fournier, A., and Hulot, G., Can one use Earth’s magnetic axial dipole field intensity to predict reversals?, *Geophysical Journal International*, **225** (1), 277-297, (2021)). The dissertation author was the primary investigator and author of this paper.

Chapter 4, is currently being prepared for submission for publication of the material.

Gwirtz, K., Davis, T. and Morzfeld, M., Can machine learning find precursors to reversals of Earth's magnetic dipole field? The dissertation author was the primary investigator and author of this material.

VITA

2009	B. Sc. in Mathematics, University of Kansas
2012	M. A. in Mathematics, University of Kansas
2019	M. S. in Applied Mathematics, University of Arizona
2021	Ph. D. in Earth Sciences, University of California San Diego

PUBLICATIONS

Gwirtz, K., Davis, T. and Morzfeld, M., Can machine learning find precursors to reversals of Earth's magnetic dipole field? (In preparation)

Gwirtz, K., Morzfeld, M., Kuang, W. and Tangborn, A., A testbed for geomagnetic data assimilation, *Geophysical Journal International*, **227** (3), 2180-2203, (2021)

Gwirtz, K., Morzfeld, M., Fournier, A., and Hulot, G., Can one use Earth's magnetic axial dipole field intensity to predict reversals?, *Geophysical Journal International*, **225** (1), 277-297, (2021)

Brio, M., Caputo, J.G., Gwirtz, K., Liu, J., and Maimistov, A., Scattering of a short electromagnetic pulse from a Lorentz-Duffing film: theoretical and numerical analysis, *Wave Motion*, **89**, 43-56 (2019).

ABSTRACT OF THE DISSERTATION

Computational tools for estimating and predicting the state of the geodynamo

by

Kyle Gwirtz

Doctor of Philosophy in Earth Sciences

University of California San Diego, 2021

Professor Matthias Morzfeld, Chair

The geodynamo is a dynamic process, involving convective motion in the Earth's fluid outer core, which is responsible for generating and maintaining the main magnetic field of the planet. That magnetic field changes on timescales ranging from decades to millions of years. On shorter timescales, the detailed morphology of the field can vary, while on millennial timescales, dramatic changes, such as reversals in the polarity of the axial dipole, can occur. We develop and examine tools for understanding and predicting variations in the Earth's magnetic field on each of these timescales.

For shorter timescales, we explore the growing field of geomagnetic data assimilation (GDA), in which observations of the geomagnetic field are merged with numerical geodynamo

models to estimate the dynamic state of the outer core and initialize forecasts of decadal-scale variations. We develop a proxy model, which is simpler and less computationally expensive than a numerical geodynamo, that allows us to systematically explore the challenges facing GDA through extensive numerical experiments. The outcome is a first of its kind testing environment for GDA and an accompanying first round of numerical experiments. The results lead us to propose an assimilation scheme for consideration in operational GDA systems and outline a path for future GDA development.

On longer timescales, we investigate predictions of major excursions and reversals of the magnetic field. This is done through considering a hierarchy of paleomagnetic reconstructions and numerical models ranging from simple scalar models to 3D geodynamos. We examine a set of prediction strategies using support-vector machines (SVMs), long short-term memory (LSTM) networks and a simple thresholding of the axial dipole field intensity. The results lead us to explicitly characterize differences in the way excursions and reversals occur among the hierarchy of reconstructions and models. This motivates the proposal of new criteria for identifying Earth-like models.

Chapter 1

Introduction

1.1 The geodynamo

The geodynamo is the dynamic process through which the main magnetic field of the Earth is generated by convection in the electrically-conducting fluid outer core. In simple terms, the magnetic field is maintained, in the absence of an external magnetic source, by induced electrical currents in the fluid. The resulting magnetic field is typically dominated by an axially aligned dipole but can vary on a range of spatial and temporal scales. On timescales of years to a few centuries the detailed morphology of the field can change (see, e.g., Jackson et al. 2000; Alken et al. 2021a). For example, the South Atlantic Anomaly (SAA), a region of low magnetic field intensity, has expanded and further weakened over recent decades (Finlay et al. 2020). Over longer timescales, it is known that the dipole-dominated magnetic field of the Earth has reversed polarity several times (Ogg 2012; Cande & Kent 1995; Lowrie & Kent 2004). Information on variations in the magnetic field comes from a variety of sources, including terrestrial and satellite based observations systems (Lesur et al. 2008; Finlay et al. 2020; Sabaka et al. 2020), ship sailing logs (Jackson et al. 2000), and paleomagnetic data (Panovska et al. 2019). The secular variation (SV) of the magnetic field provides one of the few windows we have into the workings of the deep

interior of the Earth. Additionally, being able to predict geomagnetic variations has important applications. For example, forecasting the evolution of the SAA is useful in the planning and deployment of satellites since spacecraft passing through the SAA are exposed to increased radiation that has repeatedly been the cause of damaged electronics and limited performance (Heirtzler et al. 2002). For these reasons, we explore tools for understanding and predicting the behavior of the geodynamo on both short (decadal) and long (millennial) timescales.

For understanding and predicting decadal-scale variations in the magnetic field, we focus on developing tools for geomagnetic data assimilation (DA). In geomagnetic DA, observations of the geomagnetic field are merged with dynamic models to estimate the dynamic state of the core and initialize forecasts. We focus on the scenario where the dynamic model is a numerical geodynamo (see, e.g., Kuang et al. 2010; Fournier et al. 2013; Aubert 2015; Sanchez et al. 2019; Minami et al. 2020) however, other models have been employed (see, e.g., Barrois et al. 2018; Bärenzung et al. 2020; Ropp et al. 2020). We present a proxy model that can be used to investigate some of the fundamental issues of geomagnetic DA and prototype new techniques. The approach of using proxy models to develop data assimilation techniques for a particular application has been widely and successfully employed in, for example, atmospheric DA (see, e.g. Anderson & Anderson 1999; Hamill et al. 2001; Anderson 2007). The purpose of the proxy model is to capture some of the fundamental challenges of geomagnetic DA in a system significantly less computationally complex than a numerical geodynamo. This allows for extensive numerical experiments which highlight some of the difficulties of geomagnetic DA and suggest candidate techniques for testing in operational geomagnetic DA systems.

On millennial timescales, we study predictions of reversals of the axial dipole component of the magnetic field. We take two approaches to making predictions. First we consider a threshold strategy: if the intensity of the axial dipole drops below a certain value, we predict a reversal will occur. Next, we investigate whether considering the recent history of intensity fluctuations in the axial dipole can improve predictions. To make predictions based off of a history of dipole intensity

we employ machine learning models known as support-vector machines (see, e.g., Cristianini et al. 2000). In the study of both approaches we use paleomagnetic reconstructions of the Virtual Axial Dipole Moment covering the last 2 Myr (PADM2M and Sint-2000, Ziegler et al. 2011; Valet et al. 2005) along with simulations from numerical models of varying complexity. The results reveal fundamental differences between the observational record and numerical simulations which need to be reconciled.

The contents of this chapter provide background and motivation for understanding the subsequent material, which concerns the details of the work outlined above. In the next section we provide background relevant to geomagnetic DA and the development of the proxy model. This is followed by a section outlining the approach of the studies concerning the prediction of reversals.

1.2 Data assimilation and its application to geomagnetism

In this section we present background information relevant to understanding and motivating the geomagnetic DA study of chapter 2. We begin with a brief introduction to DA, some of its common challenges, and the strategy of using proxy models to study it. This is followed by an outline of relevant aspects of DA in geomagnetism and its recent history.

1.2.1 A brief review of data assimilation

Data assimilation merges observations of a process with a computational model, in a Bayesian framework, to estimate the state of a system. It has been widely and successfully employed in applications such as numerical weather prediction (Bauer et al. 2015) over the last few decades. Many approaches to DA exist and they are typically constructed within the following framework. Let the "true" state of a system at a particular time be represented by the vector $\mathbf{x}^t$ of dimension N_x. In geomagnetic DA for example, this might represent the state

of the fluid flow, density perturbations and magnetic field of the outer core (the state of the geodynamo). Observations of the state of the system at a particular time are recorded in $\mathbf{y}$, a vector of dimension N_y. Note that we use the term observations, however $\mathbf{y}$ does not necessarily consist of unprocessed measurements. For example, in geomagnetic DA, $\mathbf{y}$ often consists of spherical harmonic coefficients, fitted to a large collection of measurements, to define a potential for the poloidal magnetic field near the core-mantle boundary (see section 1.2.4). The relationship between observations and the true state is modeled according to

$$\mathbf{y} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{H}$ is the observation operator and $\boldsymbol{\varepsilon}$ is the observation noise which is often assumed to be Gaussian with mean zero and covariance $\mathbf{R}$. Equation 1.1 defines a likelihood $p(\mathbf{y}|\mathbf{x}^t)$ which together with a prior distribution $p_0(\mathbf{x}^t)$ and Bayes' rule, defines a posterior distribution

$$p(\mathbf{x}^t|\mathbf{y}) \propto p_0(\mathbf{x}^t)p(\mathbf{y}|\mathbf{x}^t). \tag{1.2}$$

The objective of the various methods of DA is to approximate this posterior distribution.

The ensemble Kalman filter

The ensemble Kalman filter (EnKF) is a method of DA which combines a Monte Carlo approach with the Kalman filter (see, e.g., Evensen 2006) and it has been widely employed in geomagnetic DA systems (see, e.g., Fournier et al. 2013; Sanchez et al. 2020; Tangborn et al. 2021). In an EnKF, multiple, simultaneous runs of a numerical model are used to produce an ensemble of N_e forecasts $X^f = \{\mathbf{x}_1^f, ..., \mathbf{x}_{N_e}^f\}$, for a time which observations are to be assimilated. This *forecast ensemble* is treated as a sampling of the prior distribution $p_0(\mathbf{x}^t)$. The EnKF adjusts the forecasts according to the observations $\mathbf{y}$, their associated uncertainties, and the statistics of the forecast ensemble, in order to produce an *analysis ensemble* which is treated as a sampling

of the posterior distribution $p(\mathbf{x}^t|\mathbf{y})$. One can then use the mean of the analysis ensemble as an estimate of the true state with the variance of the ensemble indicating the uncertainty. The analysis ensemble can then be propagated forward by the numerical model and used as the forecast ensemble at the next time at which observations are available for assimilation. Figure 1.1 illustrates this process, through multiple assimilation cycles, for a system with two variables—one observed and one unobserved. Individual ensemble members (simulations of the two variable

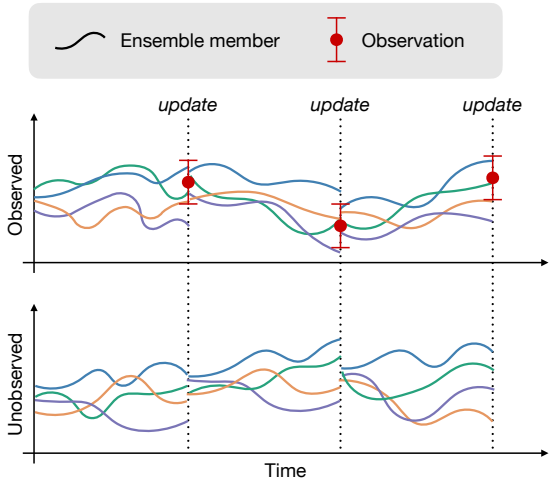

Figure 1.1: An illustration of an EnKF applied to a system of two variables. Individual ensemble members correspond to a particular color of curve. Observations (red circles) and their associated uncertainties (red bars) are used to update ensemble members.

system) correspond to a particular color of curve. When an observation is available (red dot) with an associated uncertainty (red bars) the ensemble members are updated and then propagated forward to the next assimilation time. Notice that adjustments are made not only to the observed part of the state space, but the unobserved part as well. Updates to unobserved quantities in ensemble-based DA are dependent on knowledge of correlations which are determined from the forecast ensemble. For example, in geomagnetic DA, only partial observations of the magnetic field near the core-mantle boundary (CMB) are available. Therefore, any adjustments to an estimate of the fluid flow or the unobserved magnetic field, must be made through understanding correlations between errors in estimates of the observed and unobserved components of the

geodynamo.

Many versions of the EnKF exist (see, e.g., Tippett et al. 2003; Hunt et al. 2007; Buehner et al. 2017) which differ in the details of their implementation. A common approach, known as the *stochastic EnKF*, works as follows. The ensemble of forecasts is used to compute the forecast covariance

$$\mathbf{P}^f = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\mathbf{x}_i^f - \bar{\mathbf{x}})(\mathbf{x}_i^f - \bar{\mathbf{x}})^T, \tag{1.3}$$

where $\bar{\mathbf{x}} = (1/N_e)\Sigma_{i=1}^{N_e} \mathbf{x}_i^f$. The analysis ensemble is then computed by

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}(\mathbf{y} - (\mathbf{H}\mathbf{x}_i^f + \boldsymbol{\varepsilon}_i)), \tag{1.4}$$

for $i = 1, \ldots N_e$, where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ and

$$\mathbf{K} = \mathbf{P}^f \mathbf{H}^T (\mathbf{H}\mathbf{P}^f \mathbf{H}^T + \mathbf{R})^{-1}, \tag{1.5}$$

is the estimate of the Kalman gain. All implementations of the EnKF rely on the ensemble covariance of equation (1.3) to produce a Monte Carlo approximation to the Kalman gain; and all are designed such that, under certain conditions, the analysis ensemble is distributed according to the posterior.

Variational and hybrid DA methods

Alternatives to the EnKF are variational and hybrid methods. Variational methods effectively transform the assimilation of observations into an optimization problem. Let M be a numerical model with

$$\mathbf{x}(t_k) = M[(\mathbf{x}(t_{k-1})], \tag{1.6}$$

where $\mathbf{x}(t_k)$ is the state vector at time t_k. Suppose also that we have observations $\mathbf{y}(t_k)$ for times $0 \leq k \leq n$ with Gaussian error of mean zero and covariance $\mathbf{R}(t_k)$. To simplify notation in

the equations below let $\mathbf{x}_k = \mathbf{x}(t_k)$, $\mathbf{y}_k = \mathbf{y}(t_k)$ and $\mathbf{R}(t_k) = \mathbf{R}_k$. With this notation the posterior distribution is

$$p(\mathbf{x}_0|\mathbf{y}_{0:n}) \propto p(\mathbf{x}_0) \prod_{k=0}^{n} p(\mathbf{y}_k|\mathbf{x}_k). \tag{1.7}$$

Under the assumption that the the prior and likelihood are Gaussian, finding the mean of (1.7) is equivalent to minimizing

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \boldsymbol{\mu}_0)^T \mathbf{B}_0^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_0) + \frac{1}{2}\sum_{k=0}^{n}(\mathbf{H}\mathbf{x}_k - \mathbf{y}_k)^T \mathbf{R}_k^{-1}(\mathbf{H}\mathbf{x}_k - \mathbf{y}_k) \tag{1.8}$$

where $\boldsymbol{\mu}_0$ and $\mathbf{B}_0$ are the mean and background covariance of the prior distribution. This method of variational DA is known as *4D-Var* (see, e.g., Courtier 1997). Notice that the vectors $\mathbf{x}_k$ are functions of $\mathbf{x}_0$ via the model M and therefore evaluating $J(\mathbf{x}_0)$ requires running the model, making the minimization of equation (1.8) a potentially difficult and expensive computational process. Minimizing equation (1.8) can be made easier with code for a tangent linear model $\mathbf{M}$ for the numerical model M (see, e.g., Talagrand & Courtier 1987) however constructing $\mathbf{M}$ can be a significant challenge for a large model.

In the above description of the 4D-Var method, no mention is made of how to determine $\boldsymbol{\mu}_0$ and $\mathbf{B}_0$, the mean and covariance of the prior. These may be based on the long-term statistics of the system however, it may be desirable for the prior to be flow-dependent, particularly if the above process is to be repeated over time for multiple sets of observations. For the prior mean, the resulting state estimate of the previous assimilation may be be propagated forward by M to determine $\boldsymbol{\mu}$ at the next analysis. Estimates of the prior background covariance may be propagated forward in time with the tangent linear model according to $\mathbf{B}_{k+1} = \mathbf{M}\mathbf{B}_k\mathbf{M}^T$. However, such approximations can be poor, particularly when the time over which $\mathbf{B}_0$ is being advanced is greater than the timescale of the nonlinear dynamics of M. One way of addressing this is through the use of so-called hybrid variational methods where the statistics of the prior are determined from an ensemble. For example, one may run an ensemble of 4D-Var systems (EDA; see, e.g.,

Bonavita et al. 2012) or couple an EnKF to a 4D-Var setup (E4DVar; see, e.g., Zhang & Zhang 2012), and use the ensemble covariance $\mathbf{P}^f$ for the prior.

1.2.2 Localization and inflation

Ensemble-based DA methods are constructed such that, under certain conditions, the approximations of the posterior will converge for a sufficiently large ensemble. When employing an ensemble-based DA method however, the computational expense of numerical models can limit the ensemble size at which it is reasonable to run a system. What makes an ensemble sufficiently large depends in part, on the dimension of the state space N_x (Chorin & Morzfeld 2013). If the state space is large relative to the ensemble size, sampling error can lead to spurious correlations and poor uncertainty estimates in the ensemble covariance of equation (1.3). This is an issue in geomagnetic DA where numerical geodynamos have dimensions in the millions while computational expense has typically limited DA systems to ensemble sizes in the hundreds. In other applications of DA, the challenge of small ensembles is overcome through the application of techniques known as *localization* and *inflation*. Indeed, the scientific community of atmospheric DA agrees that localization and inflation are required in all but the simplest cases (Hamill et al. 2009; Harty et al. 2021). Surprisingly, localization and inflation have only recently begun to be explored in geomagnetic DA (Sanchez et al. 2019, 2020).

Localization is designed to reduce or eliminate spurious correlations in the covariance of equation (1.3). One way this is accomplished is by taking the element-wise product of a localization matrix $\mathbf{L}$, containing elements between zero and one, and the ensemble covariance $\mathbf{P}^f$, to produce a "localized" covariance

$$\mathbf{P}^f_{\text{loc}} = \mathbf{L} \circ \mathbf{P}^f, \tag{1.9}$$

for use in the assimilation system. Typically $\mathbf{L}$ is designed such that correlations over long

distances are dampened. For example, a localization $\mathbf{L}$ might take the form

$$L_{i,j} = \exp(-(d_{i,j}/\rho)^2) \tag{1.10}$$

where $d_{i,j}$ measures the distance between the ith and jth state variable and ρ is a parameter reflecting the correlation length scale. An illustration of such a pattern of localization factors is shown in figure 1.2. The coloring corresponds to the entries of $\mathbf{L}$ which would be applied to

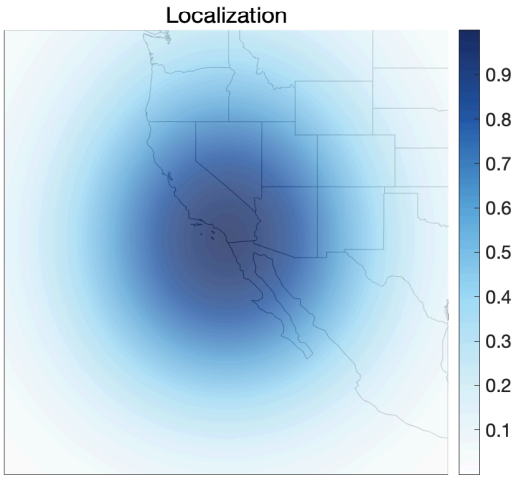

Figure 1.2: An illustration of a Gaussian decay of correlations.

covariances of $\mathbf{P}^f$ between state variables associated with La Jolla, CA (where the localization factor is one), and state variables associated with the surrounding area. While many approaches to localization exist, most rely on some measure of physical distance to dampen correlations (see, e.g., Shlyaeva et al. 2019). Unfortunately, in geomagnetic DA, such methods can be inapplicable as $\mathbf{P}^f$ often consists of covariances between spherical harmonic coefficients (see section 1.2.4) and hence, measures of physical separation are unavailable.

Inflation refers to a family of techniques which address the issue of underestimating uncertainties in a forecast ensemble (see, e.g., Kotsuki et al. 2017). For example, sampling error with a small ensemble can lead to underestimating the forecast error in $\mathbf{P}^f$, resulting in too much

weight being given to the prior and therefore limiting the impact of observations. A common approach to addressing this is *multiplicative covariance inflation* (Anderson & Anderson 1999), where, before computing $\mathbf{P}^f$, ensemble members are pushed away from the ensemble mean according to

$$\mathbf{x}^f_{i,\text{infl}} = \sqrt{1+\alpha}(\mathbf{x}^f_i - \bar{\mathbf{x}}) + \bar{\mathbf{x}}, \tag{1.11}$$

where $i = 1, ...N_e$ and $\alpha \geq 0$ is a fixed inflation parameter. Notice that this increases all entries of $\mathbf{P}^f$ by a factor of $1+\alpha$ and therefore, unlike localization, has no impact on the correlations in $\mathbf{P}^f$.

1.2.3 Developing DA with proxy models

Understanding the impact of various DA methods on a particular system calls for repeated numerical experiments. Unfortunately, for complex numerical models like those of the geodynamo, the computational expense of such experiments can be a barrier to developing DA. Additionally, interpreting the results of DA experiments with complex systems can be a challenge. For this reason, DA algorithm development and proof of concept studies often use proxy models. The purpose of a proxy model is to capture characteristics of a complex model in a simpler system of a lower computational cost, allowing for extensive numerical experiments which are easier to interpret. Two examples of models regularly used in this capacity are the Lorenz systems (Lorenz 1963, 1995) shown in figure 1.3. The Lorenz 96 system (left panel) consists of an adjustable number of coupled ordinary differential equations (ODEs) on a circle, while the Lorenz 63 system (right panel) is defined by three coupled ODEs. Both systems are inspired by the chaotic behavior of geophysical fluids and are widely used as proxy models for investigating DA (see, e.g. Anderson & Anderson 1999; Anderson 2007; Amezcua et al. 2012; Du & Shiue 2021; Kurosawa & Poterjoy 2021). Other systems which have been used as proxy models include the Kuramoto-Sivashinsky equation (Kuramoto & Tsuzuki 1975; Sivashinsky 1977), quasi-geostrophic models (see, e.g., Evensen 1994) the primitive equations (see, e.g., Ades

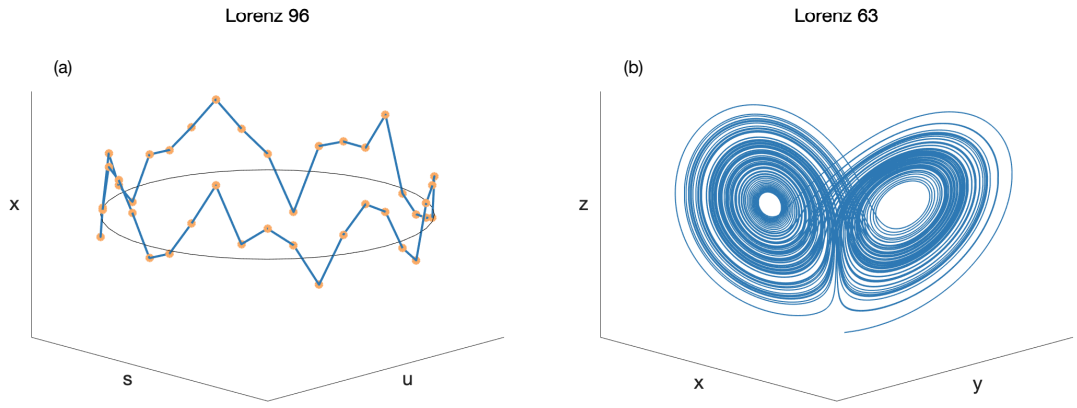

Figure 1.3: Solutions of the (a) Lorenz 96 and (b) Lorenz 63 sysetms.

& van Leeuwen 2015) the shallow water equations (see, e.g., Holm et al. 2020), and a barotropic vorticity model (see, e.g., Browne 2016). Such models have been useful for prototyping DA schemes despite not accurately reflecting all of the physics of the problems of interest. This is because results are largely intended to be interpreted qualitatively, with a proxy model primarily serving as a "gate-keeper" for determining methods which should be considered on a more complex systems. To put it simply, only DA methods shown to be robust in experiments with a proxy model should considered for further study on an operational DA system. Following this approach, we present the development of a proxy model for geomagnetic DA in chapter 2, along with a large collection of numerical experiments concerning localization and inflation.

1.2.4 Geomagnetic DA

Geodynamo models, geomagnetic observations and the challenge of localization

The magnetic field of a geodynamo model is typically decomposed into poloidal and toroidal components

$$\mathbf{B}(\mathbf{r}) = \mathbf{B}_T(\mathbf{r}) + \mathbf{B}_P(\mathbf{r}) = \nabla \times T(\mathbf{r})\hat{\mathbf{r}} + \nabla \times (\nabla \times P(\mathbf{r})\hat{\mathbf{r}}), \tag{1.12}$$

with the scalar fields $T(\mathbf{r})$ and $P(\mathbf{r})$ being further decomposed into the coefficients $T_\ell^m(r)$ and $P_\ell^m(r)$ for spherical harmonic functions of degree ℓ and order m. Other quantities, e.g., the velocity field of the outer-core are similarly defined in spherical harmonics. As a result, the state of the geodynamo is described in spherical harmonic coefficients. In the notation of section 1.2.1 this means that $\mathbf{x}$ is a vector of spherical harmonic coefficients and $\mathbf{P}^f$ consists of covariances between spherical harmonics. In geomagnetic DA, it is information on the poloidal magnetic field $P_\ell^m(r)$ near the core-mantle boundary (CMB) which is assimilated into a numerical model of the geodynamo. The observations are taken from geomagnetic field models which define magnetic field potentials in the form of spherical harmonics (see, e.g., Sabaka et al. 2020). This means that like the state of the geodynamo $\mathbf{x}$, the observations $\mathbf{y}$ are also a vector of spherical harmonic coefficients. The purpose behind assimilating spherical harmonic coefficients in this way is to isolate the signal of the geodynamo from external magnetic sources. It is believed that the large scale magnetic field, up to spherical harmonic degree $\ell = 14$, is dominated by the geodynamo (Langel & Estes 1982). Therefore, by only assimilating coefficients of degree $\ell \leq 14$, one can limit the contamination of observations by outside sources.

While the spectral nature of the observations helps isolate information on the geodynamo, it presents a major challenge to localization in geomagnetic DA. The observations and elements of the state space (all spherical harmonic coefficients) are not associated with a particular physical location and instead, are effectively global. However, localization typically makes use of measures of physical distance to limit spurious correlations from small ensembles (see section 1.2.1). Given that this approach is unavailable, alternative methods of localization are needed for geomagnetic DA. Instead of reflecting correlation length scales, a useful geomagnetic DA localization will reflect correlation structures between the spherical harmonics defining the magnetic field and the state of the fluid outer core. Exploring this issue serves as one of the primary motivations of the development and experimentation with the proxy model of chapter 2.

Current results in geomagnetic DA

Geomagnetic DA was first formally studied around fourteen years ago (Fournier et al. 2007; Liu et al. 2007; Sun et al. 2007). A number of studies have since followed which investigate the properties and dynamics of the core, study observation and model parameter errors and forecast the magnetic field (see, e.g., Kuang et al. 2009; Canet et al. 2009; Aubert & Fournier 2011; Fournier et al. 2011; Li et al. 2011; Aubert 2014; Tangborn & Kuang 2015; Sanchez et al. 2016). While geomagnetic DA is a powerful tool for exploring the Earth's deep interior, perhaps its most obvious and visible application is in the forecasting of secular variation (SV).

A widely used source of short-term SV forecasts is the International Geomagnetic Reference Field (IGRF). The IGRF is updated every five years and consists of spherical harmonic models of Earth's magnetic field over time. IGRF models are constructed from a combination of candidate models with weights determined by an international panel of scientists. Each update includes temporal derivatives for Gauss coefficients through degree eight, for the purpose of forecasting the coming five years through linear extrapolation. Just over a decade ago, the first forecast of SV through geomagnetic DA (Kuang et al. 2010) was included in IGRF-11 (Finlay et al. 2010). Since then, forecasts from geomagnetic DA systems have played an increasing role in the IGRF with the most recent release (IGRF-13 Alken et al. 2021a) including forecasts from five such systems (Bärenzung et al. 2020; Fournier et al. 2021; Tangborn et al. 2021; Minami et al. 2020; Sanchez et al. 2020). All five of these systems make use of an ensemble-based DA method (see section 1.2.1) with Minami et al. (2020) providing the first results in geomagnetic DA with a hybrid variational method. Past geomagnetic DA based forecasts contributed to IGRF models have been shown to compare favorably with other approaches to short-term (five year) predictions (Alken et al. 2021b). Beyond a few decades however, forecasts from geomagnetic DA systems have been shown to outperform linear extrapolation (Aubert 2015). This implies that current DA systems are successfully estimating and modeling some of the dynamics of the geodynamo. Current geomagnetic DA based predictions of the coming 50-100 years forecast a

continuation of the expansion and weakening of the SAA as well as the weakening of the axial dipole (Sanchez et al. 2020; Aubert 2015).

1.3 Predicting reversals of the Earth's magnetic dipole

The content of this section provides background and motivation for the studies on predicting reversals contained in chapters 3 and 4. We briefly outline the types of predictions we make, the strategies used to make them, and the value of studying predictions with a range of numerical models and observational records.

1.3.1 Reversal prediction as a yes or no question

The magnetic axial dipole field of the Earth has reversed polarity many times (Ogg 2012; Cande & Kent 1995; Lowrie & Kent 2004) over the last 150 Myr, with the most recent reversal occurring around 780 kyr in the past. We consider in chapters 3 and 4 whether such events can be predicted. Specifically, we study whether, based only on the present and recent history of dipole, one can predict whether a reversal is to occur within a specified range of time which we label the *prediction horizon* (PH). With this approach, we make only two possible predictions: yes, a reversal will occur within the PH; no; a reversal will not occur within the PH. Because a reversal can take several kyr to complete, meaningful predictions should anticipate reversals several millennia in advance. Simply put, predicting a reversal with a PH of only a few hundred years in advance is of little value considering the dipole will likely have significantly weakened and could already be described as being in the process of reversing. Forecasting geomagnetic reversals millennia in advance may seem to be an impossible task given that the limit of predictability for the magnetic field has been estimated to be around a century (Hulot et al. 2010). However, this limit concerns the the detailed structure of the field and low-frequency dynamics in the large scale features of the field may be predictable over much longer timescales. This is supported by

(Morzfeld et al. 2017) where DA with a low-dimensional model of the axial dipole demonstrated value in the prediction of reversals a few kyr in advance. We study two strategies to predicting, several millennia in advance, whether or not a reversal will occur.

1.3.2 Prediction strategies

To study predictions of reversals we first test a threshold strategy in chapter 3. The strategy is simply, when the intensity drops below a specified value (a threshold), predict a reversal. Figure 1.4 illustrates this strategy with the VADM of the paleomagnetic reconstruction PADM2M (Ziegler et al. 2011). The sign of the intensity indicates polarity with the timing of

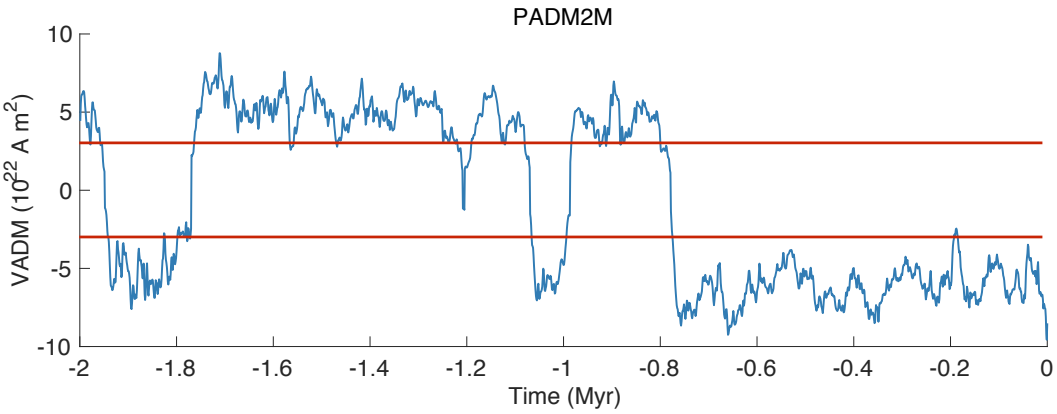

Figure 1.4: Paleomagnetic reconstruction PADM2M with sign indicating polarity and red horizontal lines indicating a threshold intensity below which one might predict a reversal.

reversals being from Cande & Kent (1995) except for a slight modification for the cobb-mountain subchron (see Morzfeld et al. 2017). The red horizontal lines indicate an intensity threshold below which one might predict a reversal is soon to occur. We vary such an intensity threshold and study its effectiveness for a collection of simulations and observational records.

In chapter 4 we apply machine learning techniques to segments of axial dipole intensity time series to make predictions. We focus on the application of support-vector machines (SVMs, see, e.g., Cristianini et al. 2000). The SVMs are trained to classify segments of time series according to whether they precede a reversal. Effectively, we investigate if SVMs can identify

signals in the dipole which indicate an oncoming reversal. Training an SVM to classify an n-dimensional object—in our case time series of dipole intensity with n data points—amounts to determining a hyperplane which separates the classes. Figure 1.5 shows how an SVM classifies a collection of test data (dots) after being trained with simulation data to predict reversals based on the dipole intensity at the present and previous time step. The present and previous intensities of

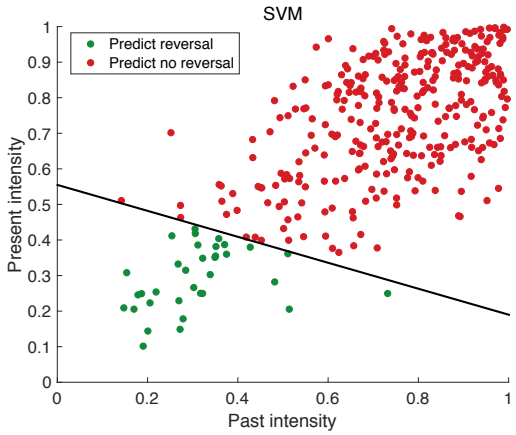

Figure 1.5: An SVM classifying data after being trained on a simulation of the magnetic axial dipole.

the test data is shown on the vertical and horizontal axis, respectively, and the values are scaled such that the long-term average intensity of the training simulation is one. The SVM classifies (predicts) whether a reversal is to occur (green) or not (red) according to which side of the black line (which is determined during training) the particular data point is. For visualization purposes we have presented an example classifying two dimensional data but in practice we consider much larger segments of dipole time series. We also test the sensitivity of results to the particular choice of machine learning model by applying long short-term memory networks (LSTMs, see, Hochreiter & Schmidhuber 1997) to the same problem.

1.3.3 Reversal predictions as a tool for characterizing dipole behavior

Because of the limited observational record of Earth's magnetic dipole, the studies of both chapters 3 and 4 make extensive use of numerical simulations in addition to paleomagnetic reconstructions. The simulations come from both low-dimensional scalar models of the dipole and a 3-D numerical dynamo. We find that the performance of the prediction strategies varies widely among the models and reconstructions. In examining the causes of these differences, the study of predictions ultimately serves as a means to characterize the behavior of the dipole and evaluate how "Earth-like" various models are. For example, the threshold strategy of chapter 3 identifies the varying ways in which models and reconstructions decay to and recover from reversals. And the machine learning study of chapter 4 highlights the relationship between the dipole's future behavior and its recent history.

1.4 Summary

This chapter is intended to present background material helpful to understaning the content and motivation of the chapters which follow. Concerning chapter 2, we discussed DA, both in general and in the context of decadal-scale variations of the geomagnetic field. The widespread use of proxy models for the study and development of new DA methods was highlighted and the challenge of developing localization strategies was emphasized as a major motivation behind the construction of a proxy model for geomagnetic DA. Other challenges to effective DA with geodynamo models remain to be explored and many of them are discussed in the details of chapter 2 and the conclusions of chapter 5.

Millennium timescale variations in the Earth's axial dipole, specifically polarity reversals, were also reviewed. We outlined two studies of the prediction of reversals. One using a simple threshold strategy to make predictions and the other, machine learning. In both cases, the strategies reveal important differences in dipole behavior among models and observational records. The

full details of these investigations can be found in chapters 3 and 4.

References

Ades, M. & van Leeuwen, P. J., 2015. The effect of the equivalent-weights particle filter on dynamical balance in a primitive equation model, *Mon. Weather Rev.*, **143**(2), 581–596.

Alken, P., Thébault, E., Beggan, C. D., Amit, H., Aubert, J., Baerenzung, J., Bondar, T. N., Brown, W. J., Califf, S., Chambodut, A., Chulliat, A., Cox, G. A., Finlay, C. C., Fournier, A., Gillet, N., Grayver, A., Hammer, M. D., Holschneider, M., Huder, L., Hulot, G., Jager, T., Kloss, C., Korte, M., Kuang, W., Kuvshinov, A., Langlais, B., Léger, J.-M., Lesur, V., Livermore, P. W., Lowes, F. J., Macmillan, S., Magnes, W., Mandea, M., Marsal, S., Matzka, J., Metman, M. C., Minami, T., Morschhauser, A., Mound, J. E., Nair, M., Nakano, S., Olsen, N., Pavón-Carrasco, F. J., Petrov, V. G., Ropp, G., Rother, M., Sabaka, T. J., Sanchez, S., Saturnino, D., Schnepf, N. R., Shen, X., Stolle, C., Tangborn, A., Tøffner-Clausen, L., Toh, H., Torta, J. M., Varner, J., Vervelidou, F., Vigneron, P., Wardinski, I., Wicht, J., Woods, A., Yang, Y., Zeren, Z., & Zhou, B., 2021a. International geomagnetic reference field: the thirteenth generation, *Earth Planets Space*, **73**(49).

Alken, P., Thébault, E., Beggan, C. D., Aubert, J., Baerenzung, J., Brown, W. J., Califf, S., Chulliat, A., Cox, G. A., Finlay, C. C., Fournier, A., Gillet, N., Hammer, M. D., Holschneider, M., Hulot, G., Korte, M., Lesur, V., Livermore, P. W., Lowes, F. J., Macmillan, S., Nair, M., Olsen, N., Ropp, G., Rother, M., Schnepf, N. R., Stolle, C., Toh, H., Vervelidou, F., Vigneron, P., & Wardinski, I., 2021b. Evaluation of candidate models for the 13th generation international geomagnetic reference field, *Earth Planets Space*, **73**(48).

Amezcua, J., Ide, K., Bishop, C. H., & Kalnay, E., 2012. Ensemble clustering in deterministic ensemble kalman filters, *Tellus A: Dynamic Meteorology and Oceanography*, **64**(1), 18039.

Anderson, J. L., 2007. Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter, *Physica D*, **230**(1), 99 – 111, Data Assimilation.

Anderson, J. L. & Anderson, S. L., 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, *Mon. Weather Rev.*, **127**(12), 2741 – 2758.

Aubert, J., 2014. Earth's core internal dynamics 1840-2010 imaged by inverse geodynamo modelling, *Geophys. J. Int.*, **197**, 1321–1334.

Aubert, J., 2015. Geomagnetic forecasts driven by thermal wind dynamics in the Earth's core, *Geophysical Journal International*, **203**(3), 1738–1751.

Aubert, J. & Fournier, A., 2011. Inferring internal properties of Earth's core dynamics and their evolution from surface observations and a numerical geodynamo model, *Nonlinear Process. Geophys.*, **18**, 657–674.

Bärenzung, J., Wicht, M. H. J., Lesur, V., & Sanchez, S., 2020. The Kalmag model as a candidate for IGRF-13, *Earth, Planets and Space*, **72**(163).

Barrois, O., Hammer, M. D., Finlay, C. C., Martin, Y., & Gillet, N., 2018. Assimilation of ground and satellite magnetic measurements: inference of core surface magnetic and velocity field changes, *Geophys. J. Int.*, **215**(1), 695–712.

Bauer, P., Thorpe, A., & Brunet, G., 2015. The quiet revolution of numerical weather prediction, *Nature*, **252**, 45–55.

Bonavita, M., Isaksen, L., & Hólm, E., 2012. On the use of EDA background error variances in the ECMWF 4D-Var, *Q. J. R. Meteorol. Soc.*., **138**(667), 1540–1559.

Browne, P. A., 2016. A comparison of the equivalent weights particle filter and the local ensemble transform Kalman filter in application to the barotropic vorticity equation, *Tellus A*, **68**, 30466.

Buehner, M., McTaggart-Cowan, R., & Heilliette, S., 2017. An ensemble Kalman filter for numerical weather prediction based on variational data assimilation: VarEnKF, *Mon. Weather Rev.*, **145**(2), 617 – 635.

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

Canet, E., Fournier, A., & Jault, D., 2009. Forward and adjoint quasi-geostrophyic models of geomagnetic secular variations, *J. Geophys. Res.*, **114**, B11101.

Chorin, A. J. & Morzfeld, M., 2013. Conditions for successful data assimilation, *J. Geophys. Res. Atmos.*, **118**(20), 11,522–11,533.

Courtier, P., 1997. Variational methods, *J. Meteorol. Soc. Japan*, **75**(1B), 211–218.

Cristianini, N., Shawe-Taylor, J., Shawe-Taylor, D., Books24x7, I., & Press, C. U., 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press.

Du, Y. J. & Shiue, M.-C., 2021. Analysis and computation of continuous data assimilation algorithms for lorenz 63 system based on nonlinear nudging techniques, *Journal of Computational and Applied Mathematics*, **386**, 113246.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res. Oceans*, **99**(10), 10143–

10162.

Evensen, G., 2006. *Data assimilation: the ensemble Kalman filter*, Springer.

Finlay, C. C., Maus, S., Beggan, C. D., Bondar, T. N., Chambodut, A., Chernova, T. A., Chulliat, A., Golovkov, V. P., Hamilton, B., Hamoudi, M., Holme, R., Hulot, G., Kuang, W., Langlais, B., Lesur, V., Lowes, F. J., Lühr, H., Macmillan, S., Mandea, M., McLean, S., Manoj, C., Menvielle, M., Michaelis, I., Olsen, N., Rauberg, J., Rother, M., Sabaka, T. J., Tangborn, A., Tøffner-Clausen, L., Thébault, E., Thomson, A. W. P., Wardinski, I., Wei, Z., & Zvereva, T. I., 2010. International Geomagnetic Reference Field: the eleventh generation, *Geophysical Journal International*, **183**(3), 1216–1230.

Finlay, C. C., Kloss, C., Olsen, N., Hammer, M. D., Tøffner-Clausen, L., Grayver, A., & Kuvshinov, A., 2020. The CHAOS-7 geomagnetic field model and observed changes in the South Atlantic Anomaly, *Earth Planets Space*, **72**(156).

Fournier, A., Eymin, C., & Alboussière, T., 2007. A case for variational geomagnetic data assimilation: insights from a one-dimensional, nonlinear, and sparsely observed MHD system, *Nonlinear Process. Geophys.*, **14**(2), 163–180.

Fournier, A., Aubert, J., & Thébault, E., 2011. Inference on core surface flow from observations and 3-D dynamo modelling, *Geophys. J. Int.*, **186**, 118–136.

Fournier, A., Nerger, L., & Aubert, J., 2013. An ensemble Kalman filter for the time-dependent analysis of the geomagnetic field, *Geochem. Geophys. Geosyst.*, **14**, 4035–4043.

Fournier, A., Aubert, J., Lesur, V., & Ropp, G., 2021. A secular variation candidate model for igrf-13 based on swarm data and ensemble inverse geodynamo modelling, *Earth Planets Space*, **73**(43).

Hamill, T., Whitaker, J., & Snyder, C., 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter, *Mon. Weather Rev.*, **129**.

Hamill, T. M., Whitaker, J. S., Anderson, J. L., & Snyder, C., 2009. Comments on "Sigma-Point Kalman Filter Data Assimilation Methods for Strongly Nonlinear Systems", *J. Atmos. Sci.*, **66**(11), 3498–3500.

Harty, T., Morzfeld, M., & Snyder, C., 2021. Eigenvector-spatial localisation, *Tellus A*, **73**(1), 1–18.

Heirtzler, J. R., Allen, J. H., & Wilkinson, D. C., 2002. Ever-present South Atlantic Anomaly damages spacecraft, *Eos, Transactions American Geophysical Union*, **83**(15), 165–169.

Hochreiter, S. & Schmidhuber, J., 1997. Long Short-Term Memory, *Neural Computation*, **9**(8), 1735–1780.

Holm, H. H., Sætra, M. L., & van Leeuwen, P. J., 2020. Massively parallel implicit equal-weights particle filter for ocean drift trajectory forecasting, *J. Comput. Phys.*, **6**, 100053.

Hulot, G., Finlay, C. C., Constable, C. G., Olsen, N., & Mandea, M., 2010. The magnetic field of planet Earth, *Space Science Reviews*, **152**, 159–222.

Hunt, B. R., Kostelich, E. J., & Szunyogh, I., 2007. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, *Physica D*, **230**(1), 112–126, Data Assimilation.

Jackson, A., Jonkers, A. R. T., & Walker, M. R., 2000. Four centuries of geomagnetic secular variation from historical records, *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, **358**(1768), 957–990.

Kotsuki, S., Ota, Y., & Miyoshi, T., 2017. Adaptive covariance relaxation methods for ensemble data assimilation: experiments in the real atmosphere, *Q. J. R. Meteorol. Soc.*., **143**(705), 2001–2015.

Kuang, W., Tangborn, A., Wei, Z., & Sabaka, T. J., 2009. Constraining a numerical geodynamo model with 100 years of surface observations, *Geophys. J. Int.*, pp. doi:10.1111/j.1365–246X.2009.04376.x.

Kuang, W., Wei, Z., Holme, R., & Tangborn, A., 2010. Prediction of geomagnetic field with data assimilation: a candidate secular variation model for IGRF-11, *Earth Planets Space*, **62**(7).

Kuramoto, Y. & Tsuzuki, T., 1975. On the formation of dissipative structures in reaction-diffusion systems, *Prog. Theor. Phys.*, **54**(3), 687–699.

Kurosawa, K. & Poterjoy, J., 2021. Data assimilation challenges posed by nonlinear operators: A comparative study of ensemble and variational filters and smoothers, *Monthly Weather Review*, **149**(7), 2369 – 2389.

Langel, R. A. & Estes, R. H., 1982. A geomagnetic field spectrum, *Geophys. Res. Lett.*, **9**(4), 250–253.

Lesur, V., Wardinski, I., Rother, M., & Mandea, M., 2008. GRIMM: The GFZ Reference Internal Magnetic Model based on vector satellite and observatory data, *Geophysical Journal International*, **173**(2), 382–394.

Li, K., Jackson, A., & Livermore, P. W., 2011. Variational data assimilation for the initial-value dynamo problem, *Phys. Rev. E*, **84**, 056321.

Liu, D., Tangborn, A., & Kuang, W., 2007. Observing system simulation experiments in geomagnetic data assimilation, *J. Geophys. Res.*, **112**, doi:10.1029/2006JB004691.

Lorenz, E. N., 1963. Deterministic nonperiodic flow, *J. Atmos. Sci.*, **20**(2), 130 – 141.

Lorenz, E. N., 1995. Predictability: a problem partly solved., *ECMWF Seminar Proceedings on Predictability*, **1**, 118.

Lowrie, W. & Kent, D., 2004. Geomagnetic polarity time scale and reversal frequency regimes, *Timescales of the paleomagnetic field*, **145**, 117–129.

Minami, T., Nakano, S., Lesur, V., Takahashi, F., Matsushima, M., Shimizu, H., Nakashima, R., Taniguchi, H., & Toh, H., 2020. A candidate secular variation model for IGRF-13 based on MHD dynamo simulation and 4DEnVar data assimilation, *Earth, Planets Space*, **72**(136).

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.

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

Panovska, S., Korte, M., & Constable, C. G., 2019. One hundred thousand years of geomagnetic field evolution, *Reviews of Geophysics*, **57**(4), 1289–1337.

Ropp, G., Lesur, V., Bärenzung, J., & Holschneider, M., 2020. Sequential modelling of the Earth's core magnetic field, *Earth, Planets and Space*, **72**(153).

Sabaka, T. J., Tøffner-Clausen, L., Olsen, N., & Finlay, C. C., 2020. CM6: a comprehensive geomagnetic field model derived from both CHAMP and Swarm satellite observations, *Earth, Planets and Space*, **72**(80).

Sanchez, S., Fournier, A., Aubert, J., Cosme, E., & Gallet, Y., 2016. Modelling the archaeomagnetic field under spatial constraints from dynamo simulations: a resolution analysis, *Geophys. J. Int.*, **207**(2), 983–1002.

Sanchez, S., Wicht, J., Bärenzung, J., & Holschneider, M., 2019. Sequential assimilation of geomagnetic observations: perspectives for the reconstruction and prediction of core dynamics, *Geophys. J. Int.*, **217**(2), 1434–1450.

Sanchez, S., Wicht, J., & Bärenzung, J., 2020. Predictions of the geomagnetic secular variation based on the ensemble sequential assimilation of geomagnetic field models by dynamo simulations, *Earth Planets Space*, **72**(157).

Shlyaeva, A., Whitaker, J., & Snyder, C., 2019. Model-space localization in serial ensemble filters, *J. Adv. Model. Earth Syst.*, **11**(6), 1627–1636.

Sivashinsky, G., 1977. Nonlinear analysis of hydrodynamic instability in laminar flames-I. derivation of basic equations, *Acta Astronaut.*, **4**, 1177–1206.

Sun, Z., Tangborn, A., & Kuang, W., 2007. Data assimilation in a sparsely observed one-dimensional modeled MHD system, *Nonlinear Process. Geophys.*, **14**, 181–192.

Talagrand, O. & Courtier, P., 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory, *Q. J. R. Meteorol. Soc.*., **113**(478), 1311–1328.

Tangborn, A. & Kuang, W., 2015. Geodynamo model and error parameter estimation using geomagnetic data assimilation, *Geophys. J. Int.*, **200**(1), 664–675.

Tangborn, A., Kuang, W., Sabaka, T., & Yi, C., 2021. Geomagnetic secular variation forecast using the NASA GEMS ensemble Kalman filter: A candidate SV model for IGRF-13, *Earth, Planets and Space*, **73**(47).

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., & Whitaker", J. S., 2003. Ensemble square root filters, *Mon. Weather Rev.*, **131**, 1485–1490.

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

Zhang, M. & Zhang, F., 2012. E4DVar: Coupling an ensemble Kalman filter with four-dimensional variational data assimilation in a limited-area weather prediction model, *Mon. Weather Rev.*, **140**(2), 587 – 600.

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.

Chapter 2

A testbed for geomagnetic data assimilation

K. Gwirtz*, M. Morzfeld*, W. Kuang[o], A. Tangborn[#]
* Cecil H. and Ida M. Green Institute of Geophysics and Planetary Physics,
Scripps Institution of Oceanography, University of California, San Diego;
[o] Geodesy and Geophysics Laboratory, NASA Goddard Space Flight Center;
[#] Joint Center for Earth Systems Technology, University of Maryland Baltimore County.
Published in *Geophysical Journal International*, 2021
doi: 10.1093/gji/ggab327

Abstract. Geomagnetic data assimilation merges past and present-day observations of the Earth's magnetic field with numerical geodynamo models and the results are used to initialize forecasts. We present a new "proxy model" that can be used to test, or rapidly prototype, numerical techniques for geomagnetic data assimilation. The basic idea for constructing a proxy is to capture the conceptual difficulties one encounters when assimilating observations into high-resolution, 3D geodynamo simulations, but at a much lower computational cost. The framework of using proxy models as "gate-keepers" for numerical methods that could/should be considered for more extensive testing on operational models has proven useful in numerical weather prediction, where advances in data assimilation and, hence, improved forecast skill, are at least in part enabled by the common use of a wide range of proxy models. We also present a large set of systematic data assimilation experiments with the proxy to reveal the importance of localization and inflation in geomagnetic data assimilation.

2.1 Introduction

Data assimilation (DA) combines observations of a geophysical process with numerical models to improve the model's predictions. Data assimilation has originated in numerical weather prediction (NWP) and has been used there with remarkable success over the past few decades (Bauer et al. 2015). Progress in DA in NWP can be attributed, at least in part, to the availability and wide use of "proxy models." A proxy model represents the numerical difficulties one encounters in the "real" problem, but the computational cost of a simulation with the proxy is much less than that of an actual simulation. The Lorenz problems (Lorenz 1963, 1995), which are simple ordinary differential equations, are prominent examples of proxy models in NWP. Other examples include the Kuramoto-Sivashinsky equation (Kuramoto & Tsuzuki 1975; Sivashinsky 1977), quasi-geostrophic models (see, e.g., Evensen 1994) the primitive equations (see, e.g., Ades & van Leeuwen 2015) the shallow water equations (see, e.g., Holm et al. 2020), and a barotropic vorticity model (see, e.g., Browne 2016). While these models may not accurately describe an atmospheric flow, they have been extremely useful for prototyping and testing new schemes for DA, which has in turn led to more accurate forecasts. Perhaps it is fair to say that new ideas for DA schemes can be tested (quickly) on proxy models, and only those that pass the test should be considered further. Thus, to maximize the usefulness of the proxy, it is important that it captures some of the major challenges of the "real" model and overall DA system.

The main contribution of this paper is to describe a proxy model relevant to geomagnetic DA. In geomagnetic DA, measurements of Earth's magnetic field are assimilated into dynamic models of the magnetic field and related systems. We focus on the scenario where observations are assimilated into self-consistent dynamo models, which couple the magnetic field to fluid motion in the Earth's liquid outer core. Alternatives to this approach, which rely on simpler models, have recently begun to be explored (see, e.g., Barrois et al. 2018; Bärenzung et al. 2020; Ropp et al. 2020). While geomagnetic DA is currently a very active field, many questions regarding

the details of DA are not fully understood. On a fundamental level, nearly all of geomagnetic DA uses an ensemble method (see, e.g., Sanchez et al. 2019, 2020; Sun & Kuang 2015; Fournier et al. 2013), but it has never been established that this approach is appropriate or more effective than, e.g., a variational method (Li et al. 2011, 2014). Indeed, only recently was the first "hybrid" ensemble-based variational geomagnetic DA system developed (Minami et al. 2020).

On a more nuanced, technical level, ensemble DA is known to be feasible only when *localization* and *inflation* are used (Hamill et al. 2009). Here, feasible means that the ensemble size is "small:" each ensemble member requires a simulation, so that a large ensemble size accrues a large computational cost. With a small ensemble size, sampling error in ensemble estimates of means and variances are large. In short, localization and inflation are techniques that reduce sampling error arising from a small ensemble size. Localization in particular reduces sampling error by enforcing an assumed correlation structure onto ensemble covariances and is well-understood and widely used in NWP and oceanography, where spatially decaying correlations are prevalent. Correlation structures in geomagnetic DA, however, are more complicated and many fundamental questions regarding effective localization and inflation are indeed unanswered (see Section 2.3 for more details). We believe that a proxy model can be useful for finding partial answers to these important questions. Surprisingly, some of the first papers on geomagnetic DA in fact used proxy models (Sun et al. 2007; Fournier et al. 2007), but this train of thought was discontinued, perhaps because this "first generation" of proxy models were overly simplified. Finally, the proxy we derive has pedagogical value and can be used as a (computationally) simple tool to teach geomagnetic DA to the next generation of students.

The rest of this paper is organized as follows. In Section 2.2, we first present a set of characteristics that define a useful proxy model for geomagnetic DA and then construct a proxy by coupling a chaotic flow to an induction equation. For computational reasons, we consider the model only in two-dimensional geometries (square and sphere [1]). We then provide background

[1]Throughout, we use the mathematical definition of a sphere, which refers a two-dimensional geometrical object, embedded in a three-dimensional space, i.e., the sphere is defined by a set of points that all have equal distance from

on data assimilation, localization and inflation in Section 2.3, before we showcase how to use the proxy model to study the effects of localization techniques in geomagnetic DA in Section 2.4. We finish the paper with a summary of our conclusions.

2.2 A proxy model for geomagnetic DA

2.2.1 Background on geodynamo models and geomagnetic DA

The geodynamo is the process in which the Earth's intrinsic magnetic field is generated and maintained by convective fluid motion in the Earth's liquid outer core. In particular, the system exhibits *dynamo action* if the fluid flow maintains the magnetic field in the absence of an external magnetic source. The fluid motion, or core convection, is driven by strong thermal and compositional buoyancy arising from the secular cooling and differentiation of the Earth during its entire evolutionary history. Geomagnetic observations and simulations have demonstrated that the geomagnetic field varies on vastly different time scales from several years, to millions of years. These variations, called the secular variation (SV), occur to the overall field intensity and its detailed morphology. Moreover, we know that the strength of the geomagnetic field is such that the Lorentz force is a major contributor to non-geostrophic core flow. The resulting fluid motion in Earth's outer core is turbulent and strongly coupled to the geomagnetic field. The latter two characteristics of the geodynamo – coupling between velocity and magnetic fields and chaotic behavior – must be properly represented in any useful proxy model.

A typical, magnetohydrodynamic (MHD), three-dimensional, self-consistent geodynamo *model* features the Navier-Stokes equation to describe the momentum balance of the electrically conducting fluid, a magnetic induction equation, and a thermodynamic equation which describes the energy source for driving the core convection. Using the Boussinesq approximation in a reference frame rotating at angular velocity $\mathbf{\Omega}$, the model is given by the set of partial differential

a specified point in three-dimensional space, which we take to be the origin.

equations (PDE)

$$\rho_0\left(\frac{\partial}{\partial t}+\mathbf{v}\cdot\nabla\right)\mathbf{v}+2\rho_0\boldsymbol{\Omega}\times\mathbf{v} = -\rho_0\nabla p+\rho\mathbf{g}+\rho_0\nu\nabla^2\mathbf{v}+\mathbf{J}\times\mathbf{B}, \quad (2.1)$$

$$\frac{\partial\mathbf{B}}{\partial t} = \nabla\times(\mathbf{v}\times\mathbf{B})+\eta\nabla^2\mathbf{B}, \quad (2.2)$$

$$\left(\frac{\partial}{\partial t}+\mathbf{v}\cdot\nabla\right)\rho = \kappa\nabla^2\rho, \quad (2.3)$$

where ρ is the fluid density with mean ρ_0, p is the pressure, ν is the kinematic viscosity, η is the magnetic diffusivity, $\mathbf{v}$ is the fluid velocity in the outer core, $\mathbf{B}$ is the magnetic field, $\mathbf{g}$ is gravity, κ is the thermal diffusivity, and $\mathbf{J}=(\nabla\times\mathbf{B})/\mu_0$ is the current density (Kuang & Bloxham 1999).

In geodynamo models, the magnetic field is typically decomposed into poloidal and toroidal components

$$\mathbf{B}(\mathbf{r})=\mathbf{B}_T(\mathbf{r})+\mathbf{B}_P(\mathbf{r})=\nabla\times T(\mathbf{r})\hat{\mathbf{r}}+\nabla\times(\nabla\times P(\mathbf{r})\hat{\mathbf{r}}). \quad (2.4)$$

The scalar fields $T(\mathbf{r})$ and $P(\mathbf{r})$ can be further decomposed into the coefficients $T_\ell^m(r)$ and $P_\ell^m(r)$ for spherical harmonic functions of degree ℓ and order m. The divergence free velocity field of the outer core is similarly described in spherical harmonics along with the scalar field representing the temperature perturbation. Upon discretization via spherical harmonics, the total model state can then be represented as a vector of spherical harmonic coefficients defining the magnetic field, velocity field and temperature perturbation at each radial level on the grid space of the numerical model.

In geomagnetic DA, geodynamo models are informed by observations of the poloidal magnetic field at, or near, the core-mantle boundary (CMB) (see, e.g., Sanchez et al. 2019, 2020; Tangborn & Kuang 2018; Fournier et al. 2013). The observations come from geomagnetic field models (see, e.g., Sabaka et al. 2020) in the form of spherical harmonic coefficients. Assimilating spherical harmonics derived from field models, rather than the "raw" data, is useful for isolating

the field of the geodynamo from sources external to the outer core. Indeed, it is generally agreed that large scale features, up to degree $\ell = 13$, are primarily generated in the outer core (Langel & Estes 1982). Modern, satellite-based measurements allow for regular observations that resolve the poloidal field near the outer core to this level. Geomagnetic field models covering earlier periods, provide a lower spatial and temporal resolution. For example, DA using archaeomagnetic-based representations of the field may only assimilate coefficients up to degree $\ell = 3$ (see, e.g., Sanchez et al. 2016).

2.2.2 A wishlist of properties for a useful proxy model

Based on the dynamical properties of the core state, the mathematical/numerical implementation of geodynamo simulations and of geomagnetic DA, we synthesize the following list of characteristics that a useful proxy model should exhibit:

(i) The proxy should consist of two coupled fields (velocity and magnetic fields);

(ii) the proxy should be amenable to a spectral discretization (spherical harmonics);

(iii) observations should be spherical harmonic coefficients of the (proxy) magnetic field;

(iv) the proxy should exhibit chaotic behavior.

Regarding the last point, we note that chaotic behavior of geodynamo models is implied by the Navier-Stokes equations and that chaos is strong in the geodynamo because the predictability of the Earth's magnetic field is known to be at most a century (Hulot et al. 2010).

Finally, we emphasize that a proxy model should be computationally less demanding than a geodynamo model, or else there is nothing to be gained by using the proxy. Given the complexity of a geodynamo model, which involves several millions of unknown state variables, a single simulation is already computationally demanding. Ensemble DA, which requires multiple simulations of the model per assimilation cycle, thus easily leads to peta-flop (floating-point operations) computations. This is perhaps not a critical issue for performing *one* data assimilation

in itself, but systematic studies of DA require *repeated* DA experiments. For example, there are several competing DA techniques in the (geophysical) literature, including ensemble DA (see, e.g., Evensen 2006) and variational methods (see, e.g., Courtier 1997). Computational limitations make it (nearly) impossible to test and compare both classes of DA methods in geomagnetic DA, but it is known that the type of DA algorithm used has a large impact on forecast skill.

On a more technical level, localization and inflation are critical for any useful ensemble DA system (Hamill et al. 2009). In short, localization and inflation are techniques that reduce sampling error, that arise from a small ensemble size (hundreds of ensemble members), but a small ensemble size is critical to keep the computational budget reasonable. Recently, the ideas of localization have appeared in geomagnetic DA (Sanchez et al. 2019, 2020). Yet, a systematic study of *how* these methods should be used in geomagnetic DA is missing because such studies require repeated and systematic DA experiments which cause the computational budget to explode. A proxy model that runs easily on a simple laptop computer, can be useful to answer these fundamental questions because, with access to high-performance computing (HPC), repeated and systematic DA experiments are easily within reach. In this context, we remind the reader again that some of the rapid progress in the application of DA in NWP can be attributed to the wide availability (and use) of proxy models such as the Lorenz models (Lorenz 1963, 1995), or quasi geostrophic models (see, e.g., Evensen 1994).

2.2.3 Mathematical derivation of the proxy model

We construct a proxy model that consists of a chaotic velocity field, coupled to a magnetic field via an induction equation. To keep computational requirements reasonable, we make the following two choices from the very beginning:

(i) The proxy is defined on a two-dimensional domain (square or sphere);

(ii) we substitute the Kuramoto-Sivashinsky (KS) equation for the Navier-Stokes equation.

We begin our description of the proxy within a generic 2D geometry and discuss the implications of this choice: self-sustained dynamo action is impossible in 2D so that the proxy requires an external magnetic source. We then describe the modified KS equation and the coupling of the resulting velocity and magnetic fields, and finally specify the proxy to square and spherical domains.

The proxy magnetic field

The magnetic field of the Earth is sustained because the magnetic induction arising from the convective flow compensates for the Ohmic dissipation resulting from the finite electrical conductivity of the core fluid (dynamo action). The 2D geometry we impose on the proxy model implies that the proxy cannot exhibit self sustained dynamo action (Cowling 1933). For this reason, the proxy is a 2D *magnetoconvection* system with a nontrivial, but steady, "background" field that sustains the magnetic field. Specifically, we write the total magnetic field as

$$\mathbf{B} = \mathbf{B}_0 + \mathbf{b} \tag{2.5}$$

where $\mathbf{B}_0$ is the steady background field ($\partial \mathbf{B}_0 / \partial t = \mathbf{0}$) and $\mathbf{b}$ is a two-dimensional, time-varying perturbation. The background field $\mathbf{B}_0$ will be chosen such that the perturbation field does not decay (see below). The perturbation magnetic field $\mathbf{b}$ is divergence free within the 2D domain and, thus, is defined by a magnetic vector potential that is everywhere normal to the 2D surface. Thus,

$$\mathbf{b} = \nabla \times A\hat{\mathbf{n}}, \tag{2.6}$$

where A is a scalar field and $\hat{\mathbf{n}}$ is the unit vector normal to the domain. Taking the normal component of the curl of equation (2.2) and using equations (2.5) and (2.6), along with the fact

that $\partial \mathbf{B}_0/\partial t = \mathbf{0}$ gives

$$\left[\nabla \times (\nabla \times \frac{\partial A}{\partial t}\hat{\mathbf{n}})\right]_{\perp} = \left[\nabla \times (\nabla \times (\mathbf{v} \times \mathbf{B}) + \eta \nabla^2 \mathbf{B})\right]_{\perp}, \tag{2.7}$$

where the subscript $\perp$ indicates the normal component of the bracketed vector quantity and where $\eta > 0$ is a constant.

The proxy velocity field and the modified Kuramoto-Sivashinsky equation

In a geodynamo model, the turbulent flow field of the outer core is described by the Navier-Stokes equation (2.1). We require that the velocity field of the proxy model be similarly chaotic, but need to avoid the computational difficulties that result from the Navier-Stokes equation. For this reason, we "replace" the Navier-Stokes equation by the Kuramoto-Sivashinsky (KS) equation

$$\frac{\partial \omega}{\partial t} = -\alpha \nabla^2 \omega - \beta \nabla^4 \omega - \gamma |\nabla \omega|^2, \tag{2.8}$$

where α, β and γ are positive constants (with appropriate units). The KS equation has its origins in the study of chaotic physical processes (Kuramoto & Tsuzuki 1975; Sivashinsky 1977) and has since been used as a proxy model for exploring chaotic behavior, particularly in fluid dynamics, (see, e.g., Hooper & Grimshaw (1985); Papageorgiou et al. (1990)) and data assimilation (see, e.g., Jardak et al. 2010; Chorin et al. 2013; Morzfeld et al. 2018). It can be shown that the mean value of solutions to the KS equation in this form increase (in magnitude) without bound but exert no influence on the evolution of perturbations about the mean. For this reason it is common to remove the mean so that solutions are centered at zero at all times (for further details see, e.g., Kalogirou et al. 2015; Armbruster et al. 1989).

We interpret the solution of the KS equation as the normal component of vorticity:

$$\omega = [\nabla \times \mathbf{v}]_{\perp}. \tag{2.9}$$

Defining the velocity $\mathbf{v}$ by a streamfunction ψ

$$\mathbf{v} = \nabla \times \psi \hat{\mathbf{n}}, \tag{2.10}$$

gives

$$\omega = [\nabla \times (\nabla \times \psi \hat{\mathbf{n}})]_{\perp}. \tag{2.11}$$

Thus, the vorticity ω, governed by the KS equation, defines the velocity field.

Finally, we need to couple the velocity to the magnetic field. In a geodynamo, this coupling occurs via the Lorentz force, which influences the fluid flow by acting on induced currents. The Lorentz force coupling is expressed through the term $\mathbf{J} \times \mathbf{B}$ in the Navier-Stokes equation (2.1) and we construct the coupling term of the proxy model in a similar way. Recall that the solution to the KS equation is related to the velocity through equation (2.9). This motivates the additional term $\lambda[\nabla \times (\mathbf{J} \times \mathbf{B})]_{\perp}$, where λ is a positive constant and $\mathbf{J} = (\nabla \times \mathbf{B})/\mu_0$ as before. The proxy then consists of the induction equation coupled to a modified KS equation,

$$\frac{\partial \omega}{\partial t} = -\alpha \nabla^2 \omega - \beta \nabla^4 \omega - \gamma |\nabla \omega|^2 + \lambda [\nabla \times (\mathbf{J} \times \mathbf{B})]_{\perp} \tag{2.12}$$

$$\left[\nabla \times (\nabla \times \frac{\partial A}{\partial t} \hat{\mathbf{n}}) \right]_{\perp} = \left[\nabla \times (\nabla \times (\mathbf{v} \times \mathbf{B}) + \eta \nabla^2 \mathbf{B}) \right]_{\perp} \tag{2.13}$$

To derive a non-dimensional form of the equations, we scale time by the diffusion timescale associated with the fourth-order term of the KS equation, i.e., $\bar{t} = (\beta/L^4)t$, we scale the magnetic field by a characteristic magnetic field intensity B, and distances by a typical length scale L. The dimensionless equations of (2.12) then become

$$\frac{\partial \bar{\omega}}{\partial \bar{t}} = -R_a \bar{\nabla}^2 \bar{\omega} - \bar{\nabla}^4 \bar{\omega} - \Gamma |\bar{\nabla} \bar{\omega}|^2 + \Lambda [\bar{\nabla} \times ((\bar{\nabla} \times \bar{\mathbf{B}}) \times \bar{\mathbf{B}})]_{\perp} \tag{2.14}$$

$$\left[\bar{\nabla} \times (\bar{\nabla} \times \frac{\partial \bar{A}}{\partial \bar{t}} \hat{\mathbf{n}}) \right]_{\perp} = [\bar{\nabla} \times (\bar{\nabla} \times (\bar{\mathbf{v}} \times \bar{\mathbf{B}}) + \frac{1}{R_m} \bar{\nabla}^2 \bar{\mathbf{B}})]_{\perp}, \tag{2.15}$$

where $R_a = \alpha L^2/\beta$, $\Gamma = \gamma/L^2$, $\Lambda = \lambda L^2 B^2/\mu_0\beta^2$ and $R_m = \beta/\eta L^2$; and horizontal bars indicate dimensionless quantities and operators, e.g., $\bar{\mathbf{B}} = \mathbf{B}/B$.

After non-dimensionalization, the proxy model is characterized by four dimensionless parameters (R_a, Γ, Λ and R_m) which have physical interpretations. In this context, we first recall that the diffusive term in the KS equation is of fourth order, while the second order term acts as a forcing (due to the negative sign). For this reason, R_a can be viewed as the ratio of the diffusive and the forcing/convective time scales, which is why we call this quantity the *modified Rayleigh number* of the proxy model. Moreover, R_a controls the length scales in the vorticity $\bar{\omega}$: the larger R_a is, the smaller are the features present in the solutions of $\bar{\omega}$ (and thus the velocity field $\mathbf{v}$ also exhibits smaller features). The nondimensional parameters Γ and Λ control the relative influence of the original nonlinear term of the KS equation and the Lorentz force, respectively. In the induction equation, the *magnetic Reynolds number* (R_m) is the ratio between the characteristic velocity (β/L^3), multiplied by the length scale, and the magnetic diffusivity. The value of R_m thus determines the influence of induction of the magnetic field relative to magnetic diffusion. For the remainder of this paper we work with the nondimensional proxy and for convenience drop the bar notation for dimensionless quantities.

Proxy model on the square

We now specify the generic proxy to a square domain with periodic boundary conditions. In this geometry, the normal vector is the z-direction ($\hat{\mathbf{n}} = \hat{\mathbf{z}}$) and the perturbation magnetic field $\mathbf{b}$ and the velocity field $\mathbf{v}$ have components only in the xy-plane. We chose the steady background field $\mathbf{B}_0 = B\hat{\mathbf{x}}$ to avoid uninteresting solutions that decay to zero. In general, any background field for which $\mathbf{B}_0 \neq \nabla \times A_0\hat{\mathbf{n}}$ for periodic A_0 prevents the decay (this can be derived by revisiting the definition of $\mathbf{b}$ by a vector potential in (2.6)). With $\mathbf{B}_0 = B\hat{\mathbf{x}}$, the proxy on a square geometry is

described by the following set of PDEs:

$$\frac{\partial \omega}{\partial t} = -R_a \nabla^2 \omega - \nabla^4 \omega - \Gamma |\nabla \omega|^2 + \Lambda [\nabla \times ((\nabla \times \mathbf{B}) \times \mathbf{B})]_z, \tag{2.16}$$

$$\omega = -\nabla^2 \psi, \tag{2.17}$$

$$v_x = \frac{\partial \psi}{\partial y}, \quad v_y = -\frac{\partial \psi}{\partial x}, \tag{2.18}$$

$$\frac{\partial A}{\partial t} = -(\mathbf{v} \cdot \nabla) A + \frac{1}{R_m} \nabla^2 A - v_y, \tag{2.19}$$

$$b_x = \frac{\partial A}{\partial y}, \quad b_y = -\frac{\partial A}{\partial x}, \tag{2.20}$$

where v_x and v_y are the components of $\mathbf{v}$, b_x and b_y the components of $\mathbf{b}$, and $\mathbf{B} = \mathbf{b} + \hat{\mathbf{x}}$ is the total magnetic field. Here, the del operator is $\nabla = \hat{\mathbf{x}} \frac{\partial}{\partial x} + \hat{\mathbf{y}} \frac{\partial}{\partial y}$ (recall that derivatives normal to the square vanish), and the Laplacian is $\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$.

Proxy model on the sphere

If the domain of the generic proxy model is a sphere, the vector normal to the domain is in the radial direction ($\hat{\mathbf{n}} = \hat{\mathbf{r}}$) and the magnetic and velocity fields $\mathbf{b}$ and $\mathbf{v}$ have components in the polar and azimuthal directions $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\varphi}}$. The background field should be chosen to prevent uninteresting solutions that decay to zero. In general, any $\mathbf{B}_0 \neq \nabla \times A_0 \hat{\mathbf{n}}$ for which A_0 is smooth and continuous over the sphere can be used, and we pick a dipole field $\mathbf{B}_0 = B[2\cos\theta \hat{\mathbf{r}} + \sin\theta \hat{\boldsymbol{\theta}}]$, where B is the characteristic magnetic field intensity. With a dipole background field, the proxy

model on the sphere is defined by the set of PDEs

$$\frac{\partial \omega}{\partial t} = -R_a \nabla^2 \omega - \nabla^4 \omega - \Gamma |\nabla \omega|^2 + \Lambda [\nabla \times ((\nabla \times \mathbf{B}) \times \mathbf{B})]]_r, \tag{2.21}$$

$$\omega = -\nabla^2 \psi, \tag{2.22}$$

$$v_\theta = \frac{1}{\sin\theta} \frac{\partial \psi}{\partial \varphi}, \quad v_\varphi = -\frac{\partial \psi}{\partial \theta}, \tag{2.23}$$

$$\begin{aligned} -\nabla^2 \frac{\partial A}{\partial t} &= -\nabla^2 [-(\mathbf{v} \cdot \nabla) A - v_\varphi \sin\theta + \frac{1}{R_m} (\nabla^2 A + A)] \\ &\quad + \frac{2}{\sin\theta} \left[\frac{\partial}{\partial \theta} (v_\varphi \sin\theta \cos\theta) - \frac{\partial}{\partial \varphi} (v_\theta \cos\theta) \right], \end{aligned} \tag{2.24}$$

$$b_\theta = \frac{1}{\sin\theta} \frac{\partial A}{\partial \varphi}, \quad b_\varphi = -\frac{\partial A}{\partial \theta}, \tag{2.25}$$

where v_θ and v_φ are the components of $\mathbf{v}$, b_θ and b_φ the components of $\mathbf{b}$, and $\mathbf{B} = \mathbf{b} + 2\cos\theta \hat{\mathbf{r}} + \sin\theta \hat{\boldsymbol{\theta}}$. The Laplacian operator on the unit sphere takes the form $\nabla^2 f = \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} (\sin\theta \frac{\partial f}{\partial \theta}) + \frac{1}{\sin^2\theta} \frac{\partial^2 f}{\partial \varphi^2}$ (and derivatives in the radial direction vanish). The above are the nondimensional form of the proxy model and the radius is used as the characteristic length scale L during nondimensionalization.

2.2.4 Properties of the proxy

Recall the list of properties a useful proxy model should have (see Section 2.2): coupled fields, spectral discretization, spectral observations, and chaotic behavior. By construction, the proxy is composed of two interacting fields, the square and spherical geometries allow for spectral discretization (via fast Fourier transform or spherical harmonic transform) and spectral observations, and it can be anticipated that the proxy is chaotic because the KS equation is well understood to produce chaotic solutions. The coupling of the fields and the chaotic behavior, however, warrant more investigation: if the nonlinearity/chaos is relatively mild, and/or if the coupling between magnetic and velocity fields is weak, then the proxy is not relevant to geodynamo models.

Table 2.1: Parameter values for the proxy model on a square and sphere.

	R_a	R_m	Γ	Λ
Square	10^3	10^{-3}	10^{-3}	4×10^7
Sphere	10^2	10^{-2}	10^{-2}	2×10^4

It may be possible to study these questions analytically, but here we discuss these properties in the context of concrete proxies (on the square and sphere) with non-dimensional parameters listed in Table 2.1. These parameters were determined in an iterative process (trying a few options and evaluating the results), and with a limited computational budget in mind. There are good reasons, however, for why the modified Rayleigh number should be large. A linearized analysis of the KS equation (no coupling with a magnetic field) shows that all modes decay if $R_a < 4\pi^2$, on the unit square domain, and if $R_a < 2$ on the unit sphere. One can thus anticipate that a large modified Rayleigh number will lead to chaotic dynamics (which we need), and we study the chaotic behavior of the proxy in more detail below.

Discretization, spectra and timescales

For computations, we discretize the proxy by a pseudo-spectral method (see, e.g., Fornberg 1996), using Fourier series (up to wave number 15) on the square, and spherical harmonics to degree and order 25 on the sphere. Time integration is implemented with exponential time differencing (see, e.g., Cox & Matthews 2002). The maximum time steps are $\Delta t = 5 \times 10^{-9}$ (square) and $\Delta t = 5 \times 10^{-7}$ (sphere). We also implement dealiasing. On the square, this is done by treating non-linear terms on a 48×48 grid in physical space and truncating solutions at directional wave-number 15. On the spherical surface, dealiasing is implemented by treating non-linear terms on a grid of $N_\theta = 39$ polar angles and $N_\varphi = 77$ azimuthal angles in physical space and truncating solutions at spherical harmonic degree and order 25. We emphasize that the proxy on the square is faster to compute with than the proxy in the sphere because the fast Fourier transform (FFT) is faster than the spherical harmonic transform.

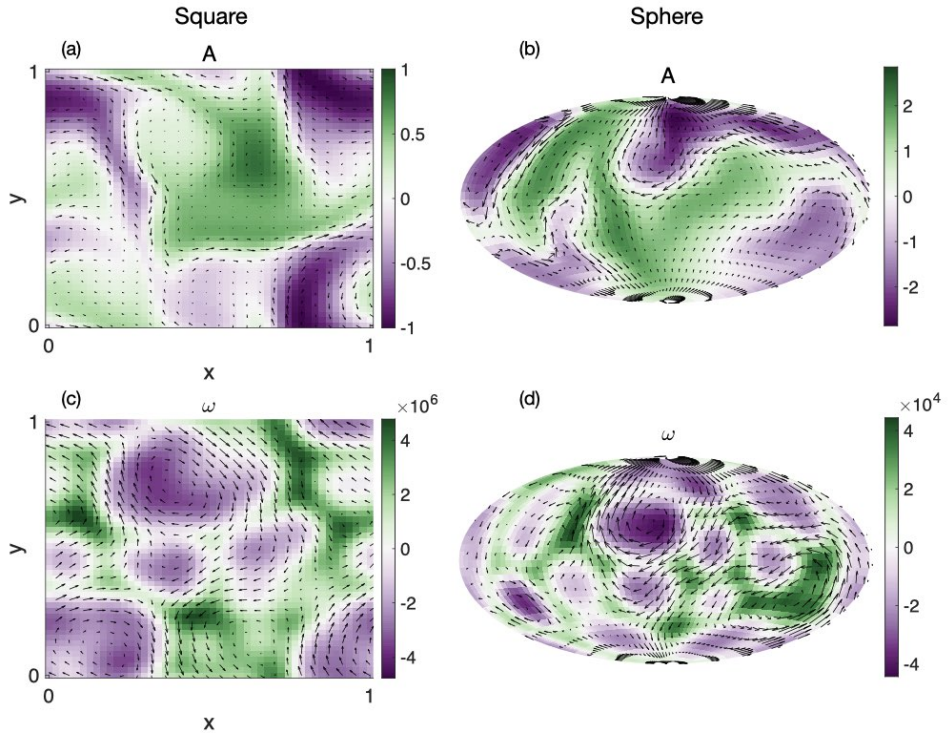

Figure 2.1: Snapshot of a solution of the proxy on the square and sphere *Top row:* Scalar defining the vector potential (color) for the perturbation magnetic field (vectors) on the (a) square and (b) sphere. *Bottom row:* Normal component of vorticity (color) and associated velocity (vectors) on the (c) square and (d) sphere.

Figure 2.1 shows a snapshot of the (numerical) solution of the proxy. The plots in the left column illustrate the proxy on the square, and the right column, the proxy on the sphere. The coloring of the top row shows the scalar A which defines the vector potential for the perturbation magnetic field, $\mathbf{b}$, depicted by the overlaid vector field. The coloring of the bottom row shows the scalar field ω, which is related to the vorticity of the overlaid velocity field. In both configurations, we note features occurring over a range of different scales, especially in the vorticity field.

We further study the power spectral density (PSD) of the proxy by generating a long simulation, and then averaging the energy in each mode (Fourier on the square, spherical harmonic on the sphere). Figure 2.2 shows the log base ten of the PSDs, scaled by the largest energy in the spectrum. We also highlight the modes that, on average, carry around 95% of the energy and note that energy is transferred between several modes, in particular in the vorticity. Nonetheless,

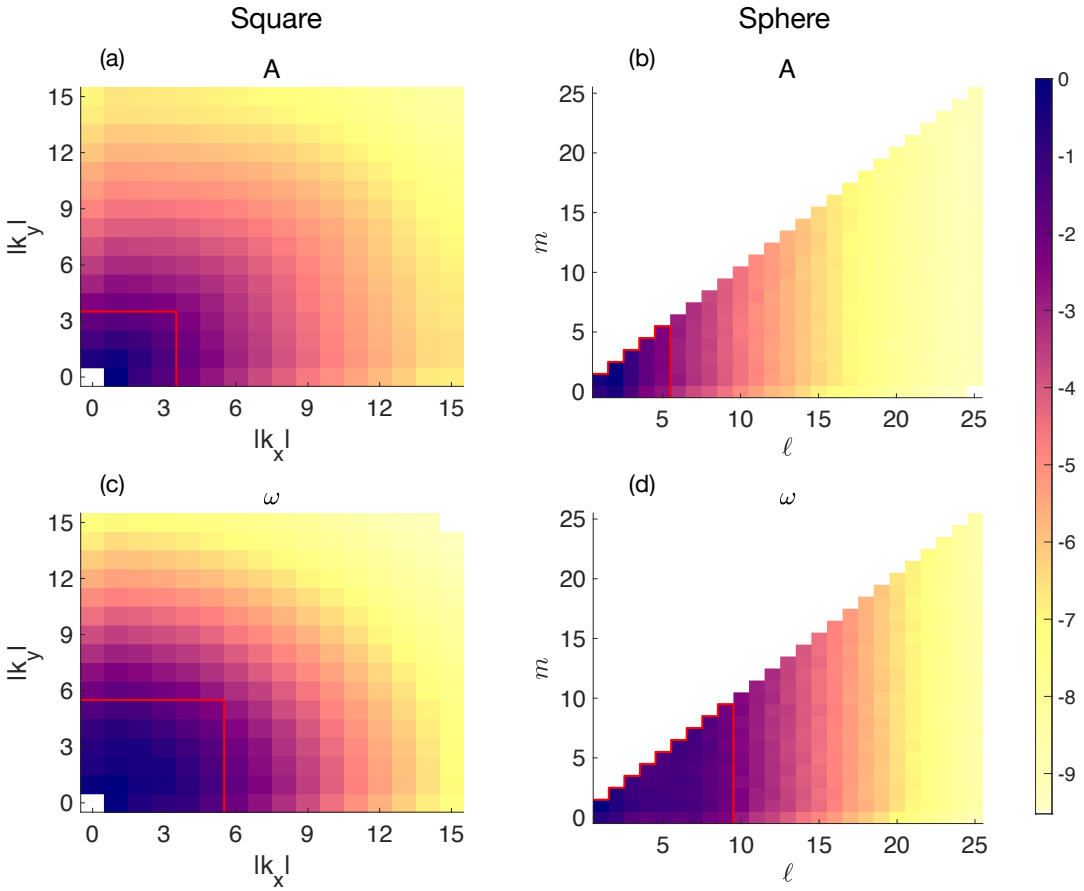

Figure 2.2: The logarithm base ten of the PSDs (scaled by the highest energy in the spectrum) of (a) A on the square, (b) A on the sphere, (c) ω on the square, (d) ω on the sphere. Red lines highlight the parts of the spectrums that carry approximately 95% of the energy, on average.

the number of modes that carry significant energy is much less than in the geodynamo or 3D geodynamo models. This is to be expected, however, because the proxy is designed to be computationally cheaper than the "real" model.

Finally, we compute typical timescales of the modes describing the fields of A and ω. For a long run on the square domain, we record for both fields, the average values of

$$\tau_{|k_x|,|k_y|} = \sqrt{\frac{\sum_{m=|k_x|} \sum_{n=|k_y|} c_{m,n}^2}{\sum_{m=|k_x|} \sum_{n=|k_y|} \dot{c}_{m,n}^2}} \tag{2.26}$$

where $c_{m,n}$ is the Fourier coefficient for wave numbers m and n in the x and y directions, respectively, and $\dot{c}_{m,n}$ is the coefficient's derivative with respect to time. Similarly, for a long run on the spherical surface domain, we record for both A and ω, the average values of

$$\tau_{\ell,m} = \sqrt{\frac{\sum_{|n|=m} (c_\ell^n)^2}{\sum_{|n|=m} (\dot{c}_\ell^n)^2}} \tag{2.27}$$

where c_ℓ^n is the spherical harmonic coefficient for degree ℓ and order n and $\dot{c}_\ell^n$ is the coefficient's derivative with respect to time. The log base ten of the resulting values are shown in Figure 2.3. We see that in both settings, the timescales vary by around an order of magnitude with the longer timescales occurring for large wavelengths. The shortest timescales on the square are a little less than 10^{-6} dimensionless time units, while the shortest timescales on the spherical surface are a little less than 10^{-4} dimensionless time units. In both settings, these timescales are over two orders of magnitude greater than the maximum allowed timesteps of $\Delta t = 5 \times 10^{-9}$ (square) and $\Delta t = 5 \times 10^{-7}$ (sphere) indicating that the time steps we chose are small enough to resolve all relevant time scales.

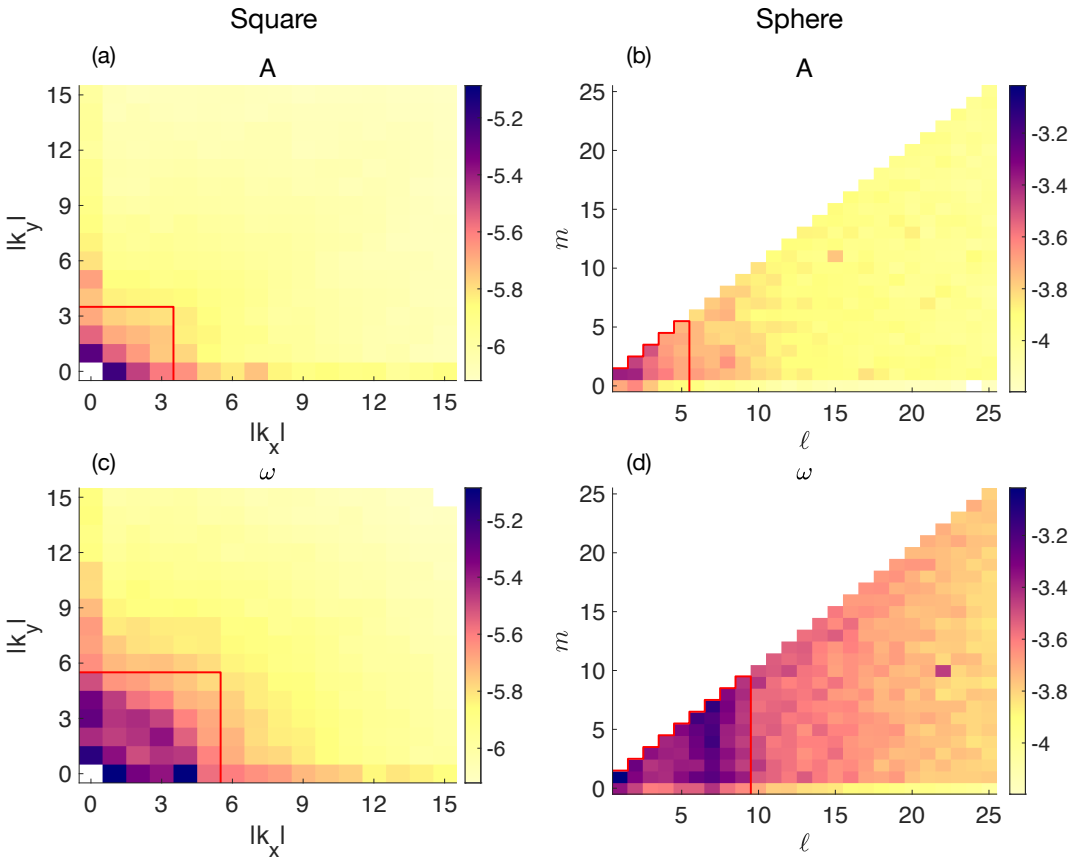

Figure 2.3: Log base ten of the typical timescales by mode of (a) A on the square, (b) A on the sphere, (c) ω on the square, (d) ω on the sphere. Red lines highlight the parts of the spectrums that carry approximately 95% of the energy, on average.

Chaotic behavior and e-folding time

We examine the sensitivity of the proxy model to initial conditions and compute the e-folding time to confirm that the proxy is indeed chaotic. To do this, we choose a set of initial conditions and generate a long run, recording the solutions at regular intervals. The same initial conditions, with a small perturbation of A, are used to generate another set of runs. In the square, the perturbation of A results from multiplying Fourier modes with directional wave numbers less than or equal to three ($|k_x|, |k_y| \leq 3$) by factors drawn from $\mathcal{N}(1, 10^{-3})$. Similarly, the initial condition of the proxy on a sphere is perturbed by multiplying spherical harmonics of A, of degree less than or equal to five ($\ell \leq 5$) by factors drawn from $\mathcal{N}(1, 10^{-3})$. For both A and ω the error is computed as the L^2 norm of the difference in the solutions, divided by the L^2 norm of the solution using unperturbed initial conditions. That is, if A and ω represent the solution at a given time using the original initial conditions and A' and ω', a solution resulting from perturbed initial conditions, we compute the errors in A and ω to be

$$\varepsilon_A = \frac{||A - A'||_2}{||A||_2}, \quad \varepsilon_\omega = \frac{||\omega - \omega'||_2}{||\omega||_2}. \tag{2.28}$$

We repeat this process for ten independent perturbations. Figure 2.4 shows the average error as a function of time; the dashed lines illustrate the log-linear fit to the average error during the period of exponential error growth. Since error growth is exponential, we conclude that the proxy is chaotic. The e-folding times—the time needed for an error to grow by a factor of e—are around 6×10^{-6} (square) and 5×10^{-4} (sphere).

Coupling of magnetic field and velocity field

The equations of the proxy model are coupled, in the sense that the magnetic field $\mathbf{B}$ appears in the equation for ω (which determines the velocity field), and, similarly, the velocity $\mathbf{v}$ appears in the induction equation which governs the evolution of the magnetic field. For a

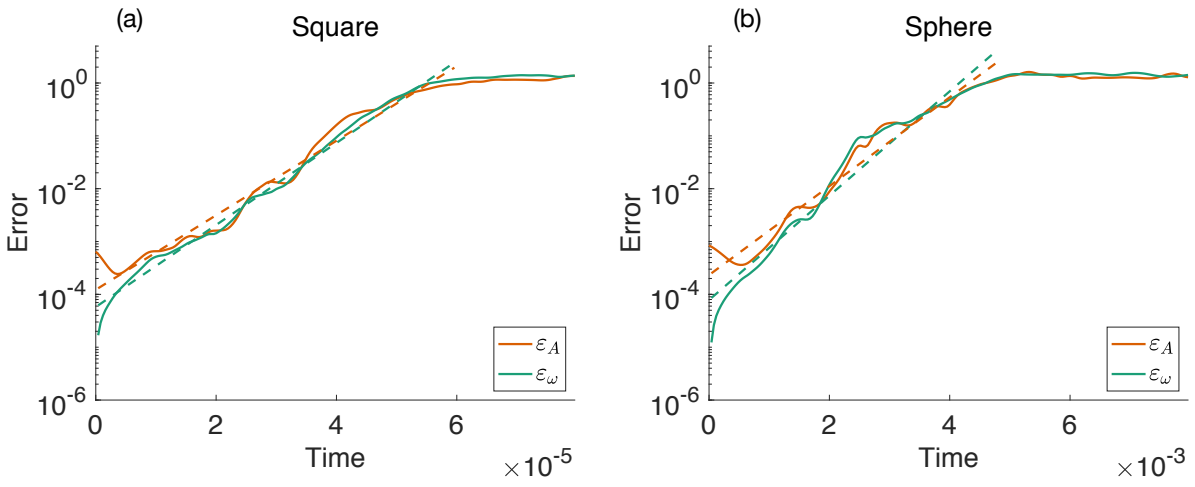

Figure 2.4: Average error vs. time for A (orange) and ω (green) on the (a) square and (b) sphere. Dashed lines show the log-linear fit to the average errors during periods of exponential growth. The slope of the dashed lines are inversely proportional to the e-folding times of the proxies on the square and sphere.

useful proxy, however, this coupling needs to be strong in the sense that the coupling terms exert a non-negligible influence on the dynamics relative to other components of the proxy. If, on the other hand, the proxy sits in a regime where the Lorentz force largely dictates the behavior of the velocity field, we risk making the challenge of estimating the velocity from magnetic field observations during data assimilation, too easy. Specifically, even if correlations between the velocity and magnetic fields are poorly understood, accurate estimates of the magnetic field could, over time, force estimates of the velocity towards its true state if the coupling is too strong.

We first consider the influence of the Lorentz force term on the velocity field. For that purpose, we compute the ratios of the L^2 norms of the forcing term ($-R_a\nabla^2\omega$) and the Lorentz force term ($\Lambda[\bar{\nabla}\times((\bar{\nabla}\times\bar{\mathbf{B}})\times\bar{\mathbf{B}})]_\perp$), over the course of a long run. Table 2.2 shows the average and standard deviation of this ratio. The L^2 norm of the Lorentz force is typically around 25% of that of the linear forcing of the KS equation in the square geometry and around 11% on the spherical surface. Second, we consider the relative influence of the velocity field on the magnetic field through induction. Similar to the comparison of forces in the modified KS equation, we compute the ratio of the L^2 norms of the induction terms and diffusion terms in the evolution

Table 2.2: The ratios of the L^2 norms of the Lorentz force and linear forcing, and the induction and magnetic diffusion terms. Standard deviations in the ratios are show in brackets.

	$\frac{\|\|\text{Lorentz force}\|\|_2}{\|\|R_a\nabla^2\omega\|\|_2}$	$\frac{\|\|\text{Induction}\|\|_2}{\|\|\text{Magnetic diffusion}\|\|_2}$
Square	0.25 (0.20)	2.13 (0.45)
Sphere	0.11 (0.06)	3.54 (1.22)

equation of A. The mean and standard deviation of this ratio is listed in Table 2.2 We find that the norm of the induction terms is typically around 2-3 times that of the magnetic diffusion, implying that the velocity field exerts meaningful influence on the magnetic field.

2.2.5 Summary of the proxy model

With geomagnetic DA in mind, we constructed a proxy model that couples a chaotic flow to an induction equation. The model exhibits chaotic behavior and meaningful coupling between the proxy magnetic and velocity fields, is amenable to spectral discretization (spherical harmonics), and runs easily on a 2021 laptop. The proxy has no natural mechanism for dipole reversals and, for that reason, should be used in the context of geomagnetic DA systems for decadal forecasts.

2.3 Using the proxy to study geomagnetic ensemble DA

By construction, the proxy is not as complex as a 3D dynamo model, but is designed to represent the major challenges in geomagnetic DA. In this section, we describe the required background on DA to use the proxy for informative numerical experiments (with results in Section 2.4). We focus on ensemble DA, because this approach is most common thus far. Within ensemble DA, localization and inflation are required numerical "tricks" that have recently been shown to have a great impact in geomagnetic DA (Sanchez et al. 2019, 2020). Localization is indeed the focus of the rest of this paper, because it is a necessary – but understudied – requirement

of computationally efficient ensemble DA (in any application of DA, including geomagnetic DA). After providing some background, we present a variety of localization schemes that are reasonable in the context of geomagnetic DA. We then put these schemes to the test in numerical experiments (Section 2.4). Being able to perform comprehensive and controlled numerical experiments in this way was a major motivation for us to create the proxy because, besides the proxy's deficiencies induced by simplifications, a detailed comparative study is out of reach with (realistic) 3D geodynamo models. Nonetheless, we emphasize that our results are qualitative and should be interpreted with the simplifications of the proxy in mind.

Finally, we understand that it is possible to use the proxy and similar numerical experiments to study if a variational approach can be more effective than ensemble DA in geomagnetic applications. We leave these ideas for the future, in part because a variational technique requires a tangent-linear and adjoint model (for efficient gradient computations), which is not currently available for any of the existing geomagnetic DA codes.

The rest of this section presents a rapid review of DA to set up notation and to introduce and explain the importance of localization and inflation. We then introduce five localization schemes that are intuitive in the context of geomagnetic DA (and in the absence of a spatial decay of correlations).

2.3.1 Review of data assimilation and the ensemble Kalman filter

DA is typically formulated within a Bayesian framework as follows. Let $\mathbf{x}$ be the vector representing the state of the system (spherical harmonic coefficients of the proxy or geodynamo), and collect all observations into a vector $\mathbf{y}$ (spherical harmonics of field models). Throughout, we use N with a subscript to denote dimensions, e.g., N_x is the dimension of the vector $\mathbf{x}$. Since the number of observations is typically much smaller than the number of state variables, we have $N_y \ll N_x$. Because the observations and model state are spherical harmonic coefficients, one can construct an $N_y \times N_x$ matrix $\mathbf{H}$, consisting of rows of the identity matrix, such that $\mathbf{Hx}$ are the

observed coefficients in $\mathbf{y}$. The state $\mathbf{x}$ and observations $\mathbf{y}$ are thus related by a linear equation

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \tag{2.29}$$

where $\boldsymbol{\varepsilon}$ is an N_y-dimensional random vector (Gaussian with mean zero and covariance $\mathbf{R}$) that models uncertainties in the model and observations. Equation (2.29) defines the likelihood $p_l(\mathbf{y}|\mathbf{x})$ which describes the probability distribution of the observations given a system state. In a sequential DA system, numerical simulations and previous assimilation results are used to define a prior distribution $p_0(\mathbf{x})$ (see below for details). By Bayes' rule, the likelihood and prior define the posterior distribution

$$p(\mathbf{x}|\mathbf{y}) \propto p_0(\mathbf{x}) p_l(\mathbf{y}|\mathbf{x}). \tag{2.30}$$

The ultimate objective of DA is the approximation of this posterior distribution. The various numerical methods used in DA differ in how approximations to the posterior distribution are found. Here, one distinguishes broadly between variational methods and ensemble methods. In variational methods, one finds the state of highest posterior probability via optimization (see, e.g., Courtier 1997). Ensemble methods, in particular the ensemble Kalman filter (EnKF) and its many variants, rely on Monte Carlo (see, e.g., Evensen 2006). Recently, hybrid techniques have been created to combine varational ideas with Monte Carlo and the ensemble approach (EDA; see, e.g., Bonavita et al. 2012; Zhang & Zhang 2012), but we do not pursue these ideas here further.

For the rest of this paper, we consider and study the EnKF, as an example of an ensemble DA technique which is the most widely used approach in geomagnetic DA. The EnKF approximates the posterior distribution by combining a Monte Carlo approach with the Kalman filter. Specifically, a *forecast ensemble* of N_e unique forecasts $X^f = \{\mathbf{x}_1^f, ..., \mathbf{x}_{N_e}^f\}$ represents a set of samples of the prior distribution $p_0(\mathbf{x})$. The aim of the EnKF is to adjust these forecasts by merging them with information contained in the observations $\mathbf{y}$. This collection of "adjusted" forecasts forms an *analysis ensemble* which is approximately distributed according to the posterior

distribution. Typically, the mean of the analysis ensemble is used as an estimate of the true state of the system, with the ensemble variance indicating the estimate's uncertainty.

One assimilation cycle of the EnKF can be implemented as follows (for other implementations, see, e.g., Tippett et al. (2003); Hunt et al. (2007); Buehner et al. (2017)). The forecast ensemble is generated by running repeatedly the (proxy) model with varying initial conditions (usually starting from the analysis ensemble of the previous assimilation cycle). The forecast ensemble defines the *forecast covariance*

$$\mathbf{P}^f = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\mathbf{x}_i^f - \bar{\mathbf{x}})(\mathbf{x}_i^f - \bar{\mathbf{x}})^T, \tag{2.31}$$

where $\bar{\mathbf{x}} = (1/N_e)\Sigma_{i=1}^{N_e} \mathbf{x}_i^f$ is the ensemble mean (or forecast mean). The *analysis ensemble* is obtained by

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}(\mathbf{y} + \boldsymbol{\varepsilon}_i - \mathbf{H}\mathbf{x}_i^f), \tag{2.32}$$

for $i = 1, \ldots N_e$, where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ and

$$\mathbf{K} = \mathbf{P}^f \mathbf{H}^T (\mathbf{H}\mathbf{P}^f \mathbf{H}^T + \mathbf{R})^{-1}, \tag{2.33}$$

is a Monte Carlo estimate of the Kalman gain.

So far, we have not specified what an appropriate ensemble size, N_e, could be. We know that a "large" ensemble size leads to small error, because Monte Carlo error reduces as $1/\sqrt{N_e}$, but it is unclear what constitutes a sufficiently "large" ensemble. Indeed, the required ensemble size depends, among other things, on the dimension of the state space $\mathbf{x}$ and the quality and extent of the observations $\mathbf{y}$ (Chorin & Morzfeld 2013). Perhaps most importantly, the computational expense of the numerical model limits the ensemble size. In geomagnetic DA, a practical ensemble size is a few hundred, which is small compared to the state dimension, which is in the millions.

2.3.2 Keeping the ensemble size manageable: Localization and inflation

The small ensemble size used in ensemble DA implies that sampling error in the Monte Carlo estimates of the EnKF is large. *Localization* and *inflation* are techniques that reduce sampling error and therefore make ensemble DA feasible at reasonable ensemble sizes. Indeed, it is widely accepted that localization and inflation are required in all but the simplest cases (Hamill et al. 2009; Morzfeld et al. 2017; Harty et al. 2021).

Localization

The idea of localization originated in NWP, where correlations are known to decay with spatial distance (hence the name *localization*). The basic idea is that sampling error causes spurious long-range correlations which should be dampened. Localization is the process in which spurious correlations and, therefore, sampling error are reduced. In geomagnetic DA, where observations are not spatial, but spectral, similar ideas can be used. The reference to a spatial correlation structure, however, must be replaced by a spectral correlation structure. Nonetheless, we will use the common terminology of localization, with the understanding that this is a misnomer and we really mean to reduce sampling error in spectral (forecast) covariances.

One common way of implementing localization is to define a localization matrix $\mathbf{L}$, and to define a *localized* forecast covariance by the element-wise product

$$\mathbf{P}^f_{\text{loc}} = \mathbf{L} \circ \mathbf{P}^f. \tag{2.34}$$

The localized forecast covariance then replaces the forecast covariance in (2.33) for the computation of the Kalman gain within an EnKF. A common choice for the localization matrix is

$$L_{i,j} = \exp(-(\mathrm{d}_{i,j}/\rho)^2), \tag{2.35}$$

where $L_{i,j}$ are the elements of $\mathbf{L}$, $\mathrm{d}_{i,j}$ is the spatial distance between the ith and jth state variables

and ρ is a length scale which is tuned (see below). Recall that the forecast covariance has (at most) rank $N_e - 1$, but if the localization matrix is high/full rank, the rank of the localized matrix is much larger than N_e (by the Schur product theorem). For this reason, a localized EnKF with a small ensemble size achieves an accuracy that far exceeds what one should expect, given the small ensemble size.

Most localization techniques in use in NWP or oceanography rely in one way or another on a measure of spatial distance between states and/or observations (see, e.g., Shlyaeva et al. 2019). All these approaches are problematic in geomagnetic DA, where no natural length scales of spatial distance exist due to the spectral nature of the state and observations. Put differently, the typical motivation for localization is to keep the effect of observations "local" (each observations only informs a subset of nearby state variables), but observations in geomagnetic DA are global (spherical harmonic coefficients). In Section 2.3.3, we describe a few localization schemes that are applicable to global/spectral observations, as they appear in geomagnetic DA.

Inflation

Sampling error due to small ensemble size results in underestimating forecast errors. A collection of methods known as *inflation* (see, e.g., Kotsuki et al. 2017) is used to compensate for this and, like localization, improve ensemble-based DA performance when ensemble size is small. In the numerical experiments below we use *multiplicative covariance inflation* (Anderson & Anderson 1999), where all entries of the forecast covariance matrix are increased by a fixed factor $\alpha \geq 0$. This is implemented by inflating the forecast ensemble

$$\mathbf{x}^f_{i,\text{infl}} = \sqrt{1+\alpha}(\mathbf{x}^f_i - \bar{\mathbf{x}}) + \bar{\mathbf{x}}, \tag{2.36}$$

for $i = 1, \ldots N_e$. The inflated ensemble has a (sample) covariance that is larger than the original sample covariance and subsequently replaces the forecast ensemble in the computation of $\mathbf{P}^f$, the

Kalman gain, and the analysis ensemble. Notice that this approach uniformly increases the entries of the forecast covariance by a factor of $1+\alpha$ and therefore, unlike localization, the adjustment has no impact on the estimated correlations of $\mathbf{P}^f$. Instead it increases the estimated uncertainty in the forecast prior with the effect that the forecast is deemed largely uncertain and, hence, more weight is given to the observations.

Tuning localization and inflation

Once one has decided on particular localization and inflation schemes, one has to determine the (hyper) parameters that define the schemes, e.g., the correlation length scale and the inflation level. This is done via a tuning process as follows. The extent of localization influences the required inflation and vice versa so that the parameters that define localization and inflation must be tuned *simultaneously*. A simple way to tune localization and inflation is to do a "grid search": A DA is performed for a pre-defined set of localization/inflation parameters and the combination that leads to the smallest errors, or best forecast skill, is deemed to be "optimal." For already computationally intensive DA systems, the computational cost of this tuning is significant and an increasing body of work is emerging that discusses approaches which require little to no experimentation to determine useful localization (see, e.g., Anderson 2012, 2016; Zhen & Zhang 2014), and inflation (see, e.g., Li et al. 2009; Gharamti et al. 2019), or both (Lunderman et al. 2021).

2.3.3 Localization schemes for geomagnetic DA

We present five localization schemes that are reasonable to try in geomagnetic DA because they do not rely on a spatial correlation structure. The first four schemes rely on a localization matrix $\mathbf{L}$ and depend on a single parameter which is tuned (see above). The first scheme, a *shrinkage* scheme, reduces cross-covariances by a uniform amount. The second scheme is based on *climatological correlations* and attempts to enforce a correlation structure on the forecast

covariance that is similar to the climatological covariance – the covariance computed from snapshots of a long run of the model. The climatological localization is most similar to the "checkerboard" localization (Sanchez et al. 2019, 2020), where the localization matrix is also the result of studying climatological correlations. The third and fourth schemes localize by *wave-vectors* or *wave-numbers*. We describe these two schemes only for the proxy on the square (Fourier modes), because it will turn out that these schemes do not perform well (see Section 2.4). For that reason, we do not perform additional testing with the spherical proxy. Nonetheless, we wish to bring up these schemes because they are somewhat natural possibilities to try, and being able to try new ideas *without* incurring a huge computational cost is one of the advantages of a proxy model. Moreover, studying how the wave-number and wave-vector localization schemes "fail" actually provides useful insights into localization in geomagnetic DA. The fifth localization scheme is the *sample error correction scheme* (SEC) (SEC, Anderson 2012). SEC differs from the above localization schemes in that it modifies the Kalman gain, rather than the forecast covariance, and that it does not require tuning because it is adaptive.

Shrinkage localization

"Shrinkage" techniques are routinely used for estimating high-dimensional covariances based on a small sample (see, e.g., Touloumis 2015). The idea is to estimate the covariance $\mathbf{P}^f_{\text{loc}}$ by a weighted sum of the sample covariance $\mathbf{P}^f$ and a target covariance $\mathbf{T}$, i.e., $\mathbf{P}^f_{\text{loc}} = (1-\rho_1)\mathbf{P}^f + \rho_1\mathbf{T}$ where $0 \leq \rho_1 \leq 1$. We use the diagonal matrix of the sample variances (the diagonal of $\mathbf{P}^f$) as the target matrix. With this choice, shrinkage localization uses a localization matrix $\mathbf{L}$ consisting of all ones along the diagonal and values of $1-\rho_1$ on all off-diagonal entries, and effectively reduces all correlations by a factor of ρ_1. Shrinkage localization is useful if there is a net benefit to reducing *all* sample-based correlations, to avoid the damaging effects of spurious correlations that arise from a small ensemble size.

Climatological localization

The rationale for climatological localization is that the forecast covariance matrix may have a similar correlation structure as the climatological covariance. One thus first performs a long model run, with initial conditions from a well developed solution, and collects "snapshots" of the model, taken at regular (or even irregular) time intervals. The snapshots are used to generate a correlation matrix that represents the climatological correlations. Subsets of these matrices are shown in figure 2.5. In the top row we see that climatological correlations between modes of the magnetic field (A) are nearly zero with the exception of some weak correlations between low-order spherical harmonics for the model on the sphere. The bottom row of figure 2.5 indicates that for both geometries there exist strong climatological correlations between modes of the magnetic field (A) and velocity field (ω). A localization matrix is obtained from the climatology, by raising the absolute values of the entries of the correlation matrix to the power $1/\rho_2$, where $\rho_2 > 0$ is a localization parameter (to be tuned). In short, the effect of a climatological localization is that the stronger the correlation between two elements of the state space over time (according to the offline, long, free model run) the more a sample correlation in $\mathbf{P}^f$ is trusted. Climatological localization is similar to the "checkerboard" localization (Sanchez et al. 2019, 2020), where a localization is constructed from climatological correlations.

Wave-vector localization

We base the localization on a "distance" defined by wave-vectors. Specifically, we retain correlations between Fourier modes of similar total wave number and orientation, but dampen correlations between modes of different length-scales or orientations. We implement wave-vector localization via the localization matrix $L_{i.j} = \exp(-(\mathrm{d}_{i,j}/\rho_3)^2)$ with $d_{i,j} = ||\mathbf{k}^i - \mathbf{k}^j||$ where $\mathbf{k}^i$ and $\mathbf{k}^j$ are the wave-number vectors of the Fourier modes corresponding to the coefficients in the ith and jth elements of the state space, and $||\cdot||$ denotes the Euclidean norm. Thus, modes of similar total wave number but different orientation are "far apart" and sample correlations

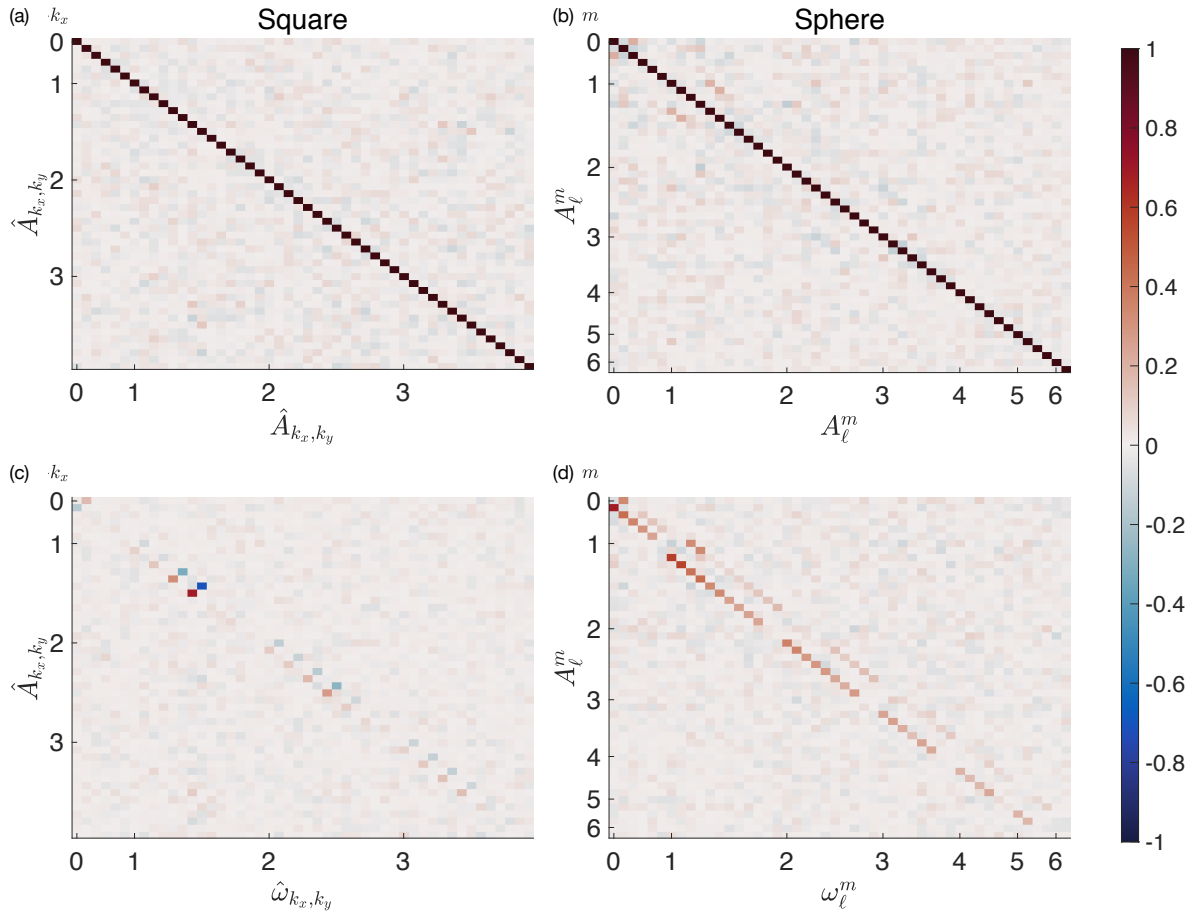

Figure 2.5: Subsets of the climatological correlation matrices of the proxy on the square (left column, grouped by k_x and truncated at $|k_x|, |k_y| = 3$) and the proxy on the sphere (right column, grouped by m and truncated at $m = 6$). (a) Correlations of the proxy on the square between modes of A. (b) Correlations of the proxy on the sphere between modes of A. (c) Correlations of the proxy on the square between modes of A and ω. (d) Correlations of the proxy on the sphere between modes of A and ω.

between them are suppressed. For example, if $\mathbf{k}^i = 3\hat{\mathbf{x}} + 3\hat{\mathbf{y}}$ and $\mathbf{k}^j = 3\hat{\mathbf{x}} - 4\hat{\mathbf{y}}$, then the modes have a similar total wave number but a different orientation and we get $d_{i,j} = 7$ (strong localization). If, on the other hand, $\mathbf{k}^i = 3\hat{\mathbf{x}} + 3\hat{\mathbf{y}}$ and $\mathbf{k}^j = 3\hat{\mathbf{x}} + 4\hat{\mathbf{y}}$, then the modes have a similar total wave number *and* orientation and, thus, $d_{i,j} = 1$ (mild localization).

Wave-number localization

Sample correlations between modes corresponding to similar total wave numbers are largely retained while sample correlations between modes of different length-scales are reduced. The localization matrix takes the form $L_{i.j} = \exp(-(\mathrm{d}_{i,j}/\rho_4)^2)$ with $d_{i,j} = |k^i - k^j|$ where k^i and k^j are the total wave numbers of the Fourier modes corresponding to the coefficients in the ith and jth elements of the state space. With this approach, only the difference in length scales of Fourier modes are relevant. For example, by this measure, elements of the state space corresponding to modes with wave-number vectors of $\mathbf{k}^i = 3\hat{\mathbf{x}} - 4\hat{\mathbf{y}}$ *or* $\mathbf{k}^j = 3\hat{\mathbf{x}} + 4\hat{\mathbf{y}}$ will be "equidistant" from other parts of the space, since for either mode the total wave number is $k = \sqrt{3^2 + 4^2} = 5$.

Sampling error correction (SEC)

In the sample error correction scheme (SEC, Anderson 2012), the larger the magnitude of a sample correlation, and the larger the ensemble from which it is computed, the more it is trusted. The SEC scheme does not operate through covariance localization as the four schemes described above. Instead, each element of the Kalman gain (see (2.33)) is multiplied by a factor between zero and one, which depends only on the ensemble size and the sample correlation between the observed state and the element of the state space being adjusted. In general, the larger the ensemble and the larger the magnitude of the sample correlation, the closer the factor is to one. An offline Monte Carlo simulation is used to construct a lookup table of factors as a function of ensemble size and sample correlation.

2.3.4 Observing system simulation experiments (OSSEs)

One can use the proxy to perform a systematic numerical study via observing system simulation experiments (OSSEs). In OSSEs (sometimes also called twin experiments), one generates observations using a model and subsequently assimilates the synthetic observations back into a model. The simulation with the model that generates the data is called the *nature run*. OSSEs have the advantage that one has full control over the observation network, observation errors, and one also has full knowledge of the system states, which allows the study of (forecast) errors in observed *and* unobserved quantities. For this reason, OSSEs are a typical first step when testing the impact of new observations, or for testing DA technology in many Earth science problems. In fact, OSSEs are routinely used at NASA and NOAA (see, e.g., Zeng et al. 2020; Hoffman & Atlas 2016; Errico et al. 2013).

The proxy models can be used for OSSEs. A long run of the models is performed using initial conditions taken from well-developed solutions (the end of long runs starting from random perturbations). A portion of these solutions are recorded as the nature runs and from the remaining part, we randomly draw initial ensemble members as needed. Choosing ensemble members from a long run reduces the risk of the initial ensemble members being correlated or impacted by a transition phase of the model. Emulating the situation in geomagnetic DA, it is reasonable to only observe the low-frequency spectral coefficients of the proxy magnetic field. In our numerical experiments below, we use Gaussian additive errors with a diagonal observation error covariance $\mathbf{R}$ (uncorrelated errors) to generate the synthetic observations. We then supply the DA system with knowledge of the error statistics, but one can easily consider different choices. In the experiments below, the nature run is generated with the proxy that is used during the assimilation, but future studies may use different configurations. For example, one can use a higher resolution proxy in the nature run, and a lower resolution proxy for the DA. One can also envision using one set of parameters (e.g., modified Rayleigh or magnetic Reynolds numbers) in the model that generates the synthetic data, and another set of parameters in the model that assimilates these

data. One can further supply the DA with false information about observation errors, e.g., one can vary the observation error covariance or one could generate synthetic observations that violate the usual Gaussian error assumptions. Such a setting may in fact more realistically represent geomagnetic DA, where coarse models are known to be biased, and where many assumptions or approximations are known to be inadequate. Nonetheless, the experiments with idealized configurations we present below is a typical and necessary first step towards understanding how localization can function (or not) in geomagnetic DA.

2.4 Results of data assimilation experiments

We carry out a series of DA experiments (OSSEs) with the proxy model. These experiments are designed to reveal if localization can help reduce the required ensemble size in EnKF and, therefore, make EnKF more effective in geomagnetic DA. We begin by discussing the general framework of the numerical experiments and metrics for evaluation of DA performance. Because the proxy on the square is easier to compute with (FFT vs. spherical harmonic transforms), we perform a large number of experiments with the proxy on the square, learn our lessons, and present a condensed set of numerical experiments with the spherical proxy. This hierarchical approach is effective because the proxies on the square and sphere share many characteristic features, but the proxy on the sphere, is dynamically more complex (more modes carry energy and more energy is transferred across spatial and temporal scales).

2.4.1 Observation network and metrics of success

All numerical experiments are OSSEs as explained above. We use the same observation network in all numerical experiments, i.e., the observed coefficients and the time interval between observations is the same for all experiments. Specifically, on the square, we observe Fourier modes of A with wave numbers in the x and y directions that are less than or equal to three. On

the sphere, we observe coefficients of A corresponding to spherical harmonics of degree five or less. In both geometries, the observed modes represent around 95% of the mean energy in the magnetic perturbation field. The time interval between observations, the *analysis time*, is 20% of the typical timescale of the magnetic perturbation field (the mean of $||A||_2/||\partial A/\partial t||_2$) or around 7% of the e-folding time. We assume (and use) Gaussian observation errors with a diagonal observation error covariance ($\mathbf{R}$). The standard deviations, which are the square roots of the diagonal elements of $\mathbf{R}$, are set to 1% of the respective coefficient's mean magnitude over the course of the nature run.

To evaluate the performance of the DA we consider the *observation-minus-forecast* (OmF) residual, the *truth-minus-forecast* (TmF) error and the *ensemble spread*. The OmF of one DA cycle is defined as

$$\text{OmF} = \sqrt{(\mathbf{H}\bar{\mathbf{x}} - \mathbf{y})^T(\mathbf{H}\bar{\mathbf{x}} - \mathbf{y})}, \tag{2.37}$$

where $\bar{\mathbf{x}}$ is the mean of the forecast ensemble (the ensemble prior to assimilating the observations $\mathbf{y}$). In other words, we take the square root of the sum of the squares of the differences in the observations and their forecasted values. Note that for the normalized Fourier and spherical harmonic coefficients, this amounts to transforming the difference into physical space and computing the L^2 norm over the domain. For this reason, we do not average the errors (divide by the number of observations) as in the root mean squared error (RMSE). Our main motivation for OmF is that is a measure of performance in an operational DA system.

TmFs (truth minus forecast) are very useful for diagnosing how well a DA system performs on observed and unobserved quantities, but since TmF requires knowledge of a "true" state, TmFs can only be used in OSSEs. We compute TmFs separately for the magnetic field and velocity field, because of their differing orders of magnitude. We thus define

$$\text{TmF}_A = \sqrt{(\bar{\mathbf{x}}_A - \mathbf{x}_A)^T(\bar{\mathbf{x}}_A - \mathbf{x}_A)}, \quad \text{TmF}_\omega = \sqrt{(\bar{\mathbf{x}}_\omega - \mathbf{x}_\omega)^T(\bar{\mathbf{x}}_\omega - \mathbf{x}_\omega)}, \tag{2.38}$$

where $\mathbf{x}_A$, $\mathbf{x}_\omega$ are from the nature run ("truth"), and $\bar{\mathbf{x}}_A$, $\bar{\mathbf{x}}_\omega$ are the forecast means of the EnKF. We note that this amounts to computing, in the physical domain, the L^2 norm of the difference between the forecast and nature run, for the fields A and ω. As with OmF, the defintion of the TmFs is, therefore, different from the usual RMSE.

The ensemble spread (spread, for short) typically indicates an average variance in the forecast and is defined by $\sqrt{\mathrm{tr}(\mathbf{P}^f)/n}$, where $\mathbf{P}^f$ is the forecast ensemble covariance matrix, $\mathrm{tr}(\mathbf{P})$ denotes the trace of a matrix $\mathbf{P}$ (the sum of the diagonal elements), and where n is the overall state dimension. The spread is compared to the forecast root mean square error (RMSE) and in a well-tuned and well-working DA system, the spread should be approximately equal to the RMSE, so that expected errors are comparable to actual errors. In geomagnetic DA, using non-averaged errors such as TmF and OmF is more reasonable because of their physical interpretation (see above). For this reason, we modify the usual definition of spread to be in line with the definitions of TmF and OmF as follows. We compute spread separately for the observed magnetic field, the overall magnetic field, and the velocity field as

$$\mathrm{Spread}_{\mathrm{obs}} = \sqrt{\mathrm{tr}(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})}, \quad \mathrm{Spread}_A = \sqrt{\mathrm{tr}(\mathbf{P}_A^f)}, \quad \mathrm{Spread}_\omega = \sqrt{\mathrm{tr}(\mathbf{P}_\omega^f)}, \tag{2.39}$$

where $\mathbf{P}_A^f$ and $\mathbf{P}_\omega^f$ are covariances of the ensemble of forecasts for the spectral coefficients of the fields A and ω. With these definitions, an indicator for a well-tuned DA system is that

$$\mathrm{OmF} \approx \mathrm{Spread}_{\mathrm{obs}}, \quad \mathrm{TmF}_A \approx \mathrm{Spread}_A, \quad \mathrm{TmF}_\omega \approx \mathrm{Spread}_\omega. \tag{2.40}$$

We note the addition of the observation error covariance ($\mathbf{R}$) in (2.39), which is necessary if observation errors are comparable to estimation errors to avoid overfitting the data.

Throughout the rest of this paper, we scale OmFs, TmFs and corresponding spreads by the macroscopic errors. This has the effect that an error of approximately one means that the DA is ineffective – it produces forecast errors about as large as what one would get with an independent

simulation. The macroscopic errors are determined from the same runs used to estimate the e-folding time (see Section 2.2.4). Specifically, we average OmF and TmFs, in time and over the perturbations, once these errors stopped to increase exponentially.

2.4.2 Results: proxy on the square

We present and discuss results of DA experiments with the proxy model on the square. As mentioned above, we perform a large number of experiments with this proxy because of its numerical simplicity and computational affordability compared to the spherical proxy (FFT is faster than spherical harmonic transform).

What error reduction can we expect: Experiments with large ensembles

Since computations with the proxy on the square are easy, we can use EnKFs with large ensemble sizes and without localization or inflation. In fact, we can determine the smallest ensemble size that leads to asymptotically small errors (large ensemble size limit), i.e., forecast errors have stabilized, and no improvements result from further increasing the ensemble size. Forecast errors in the large ensemble size limit will serve as a baseline for the localization and inflation techniques. If localization and inflation are appropriate, one should be able to obtain similarly small errors, but with a much smaller ensemble size.

The nature run for these experiments is 601 analysis times long (over 40 e-folding times) and we apply the EnKF with ensemble sizes from $N_e = 300$ to $N_e = 1000$. The mean and standard deviation of OmFs and TmFs are computed over the last 300 cycles to avoid contamination of the errors by the EnKF's spin-up period (see Section 2.4.2). We apply no localization or inflation. The results are summarized in Figure 2.6, which shows OmF and TmFs as a function of the ensemble size. The OmFs drop sharply (two orders of magnitude) and become relatively stable around $N_e = 500$, with the TmFs in the magnetic field and the unobserved velocity leveling off in the same way for ensemble sizes around $N_e = 800$. With smaller ensemble sizes, we observe

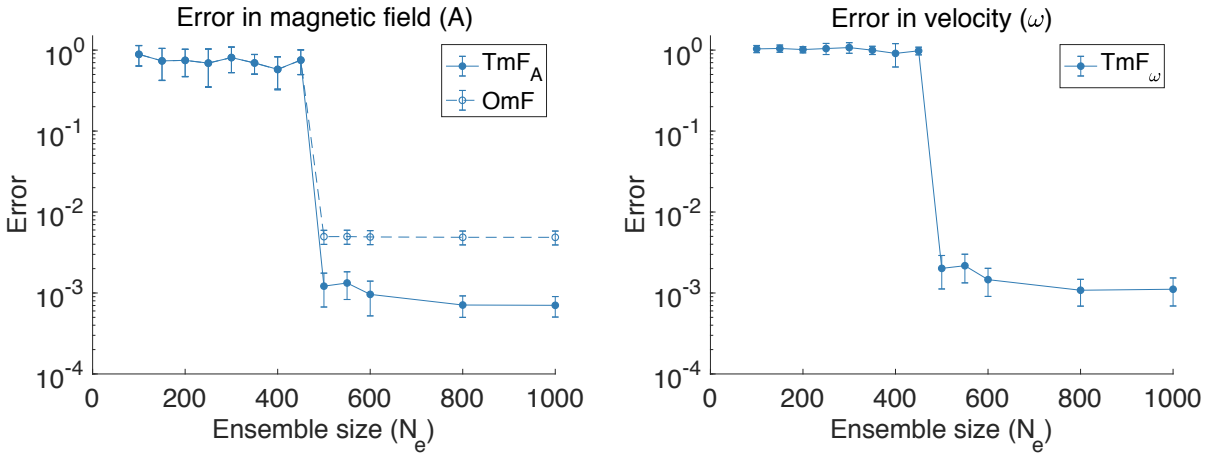

Figure 2.6: Forecast errors as a function of ensemble size (no localization or inflation). (a) Errors in magnetic field (OmF and TmF). (b) Errors in velocity (OmF). All errors are scaled by corresponding macroscopic errors, so that an error of about one means that the DA is not reducing forecast errors. The open/filled dots are average errors and the error bars indicate one standard deviation (all averages are over time).

macroscopic error, and the DA has no effect. We thus conclude that

(i) an ensemble size of at least $N_e = 500$ is required for stable, reliable and accurate DA;

(ii) the DA is able to reduce errors by more than two orders of magnitude.

It is thus clear that DA can have a huge effect on forecast accuracy, but this accuracy comes at the cost of a large ensemble size. Moreover, if the proxy we created is indeed simpler than realistic geodynamo models, then our experiments suggest that ensemble DA *without* localization or inflation requires ensemble sizes much larger than 500. For example, an ensemble size of $N_e = 1000$ is not sufficient for effective DA on the spherical proxy (see section 2.4.3), which is only mildly more dynamically rich than the proxy on the square. The remaining numerical experiments, however, are designed to investigate if localization and inflation can boost the computational efficiency of geomagnetic DA by achieving the accuracy of a large ensemble EnKF at a much smaller ensemble size.

Finally, we note that OmF can be larger than TmF. The reason is that observation errors (the additive Gaussian noise) comes into play for OmF, but not for TmF. Because the observation

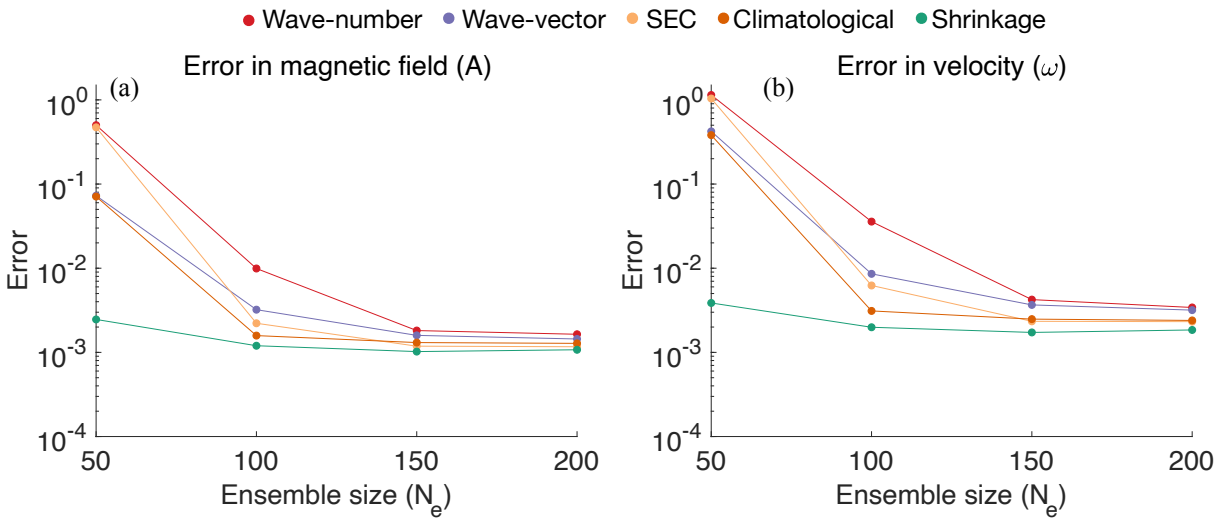

Figure 2.7: Forecast errors (TmF) as a function of ensemble size (N_e) for EnKFs when using optimal levels of localization and inflation (for each particular combination of scheme and ensemble size). (a) Errors in magnetic field. (b) Errors in velocity. All errors are scaled by corresponding macroscopic errors.

errors are larger in magnitude than the unobserved components of A we have that OmF>TmF$_A$.

Computational efficiency of DA with (tuned) localization and inflation

We now apply localization and inflation. This requires tuning via a grid-search: We chose a number of localization and inflation parameters, perform a DA experiment with each combination, and then compute the forecast error. This results in a gridded error, and we base an "optimal" localization on these errors (see Section 2.4.2 for more details). The optimal localization and inflation further depend on the ensemble size, so we repeat this tuning step at each ensemble size we consider.

With tuned localization and inflation, we carry out an OSSE with 601 analysis times and neglect the first 300 DA cycles as spin-up. The errors (TmFs) one obtains with a localized and inflated EnKF are illustrated in Figure 2.7, where we show TmFs as a function of ensemble size. The numerical experiments reveal three interesting facts:

(i) All localization schemes we tried lead to dramatically reduced errors, but at small ensemble

Table 2.3: Average errors (OmF and TmFs), scaled by the respective macroscopic errors and multiplied by 10^{-3} (proxy on the square). The large ensemble EnKF ($N_e = 1000$) does not use localization or inflation. All other EnKFs are run with ensemble size $N_e = 100$ and use the optimal levels of localization and inflation (for the particular scheme and $N_e = 100$).

		Magnetic field			Velocity	
	OmF	$\text{Spread}^f_{\text{obs}}$	TmF^f_A	Spread^f_A	TmF^f_ω	Spread^f_ω
Large Ens.	4.9	4.9	0.7	0.7	1.1	1.1
Shrinkage	5.0	5.1	1.2	1.7	2.0	3.3
Climatological	5.1	5.2	1.6	1.8	3.1	3.8
Wave-vector	5.7	5.7	3.2	3.4	8.6	9.1
Wave-number	9.4	11.8	9.9	14.0	35.8	55.3
SEC	5.3	5.4	2.2	2.4	6.2	7.4

sizes.

(ii) How quickly errors converge with ensemble size depends on the localization scheme used, with the shrinkage scheme exhibiting the fastest convergence.

(iii) The localization scheme has an effect on the smallest error one can achieve with a localized EnKF.

Thus, while every localization leads to dramatic improvements over the "vanilla" EnKF without localization and inflation, the details of the localization are indeed very important. This is perhaps intuitive because the localization scheme that most accurately reflects the "true" underlying correlation structure should lead to the smallest errors.

We now focus on errors at ensemble size $N_e = 100$ to investigate how the details of the localization affect forecast error. The errors for all five localization schemes and for the large ensemble EnKF ($N_e = 1000$) are summarized in Table 2.3. As indicated above, all localization schemes drastically reduce OmFs (by a factor of 2 or more). Indeed, the OmFs of the localized EnKFs with $N_e = 100$ are comparable to the OmFs of the EnKF with $N_e = 1000$ *without* localization and inflation (perhaps with the exception of wave-number localization). Moreover, the ensemble spread is comparable to the forecast error, indicating that the DA works as it is supposed to. Thus, when considering OmF only, it is difficult to say which localization scheme is most

appropriate, because the differences between the shrinkage, climatological, SEC and wave vector schemes are minor. Perhaps the wave-number localization can be ruled out at this stage because the OmFs are much larger than for the other schemes.

The various differences between the localization schemes are more dramatic when one considers TmFs. In fact, none of the localization schemes we tried achieves TmFs that are comparable to those of the large ensemble EnKF. This means that none of the schemes we applied accurately capture the underlying correlation structure. Nonetheless, the shrinkage scheme overall gives the smallest errors, followed by climatological localization, SEC, and wave-vector localization. Wave-number localization again leads to much larger errors. We also note that the ensemble spreads for TmFs tends to overestimate the uncertainty (Spread $>$ TmF). What this implies is that while different localization schemes may perform similarly when considering OmF, errors in unobserved system components can be dramatically different. This is important in geomagnetic DA, where one of the main motivations is to infer the core state from measurements of only the magnetic field. Our numerical experiments suggest that it is difficult to determine a "truly appropriate" localization scheme based on only short-term forecasts of observed quantities.

One way forward is to consider longer range forecasts, i.e., forecasts that stretch over more than one assimilation window. The reasons are as follows. Errors in unobserved quantities become more important as time evolves, due to nonlinear couplings. More appropriate localization schemes improve estimates of unobserved quantities, because a more appropriate correlation structure is enforced. Thus, the better the localization, the smaller the forecasts errors. This is true at any scale, but when considering observed quantities only (OmF), the differences between localization schemes become more apparent with longer range forecasts.

We now test this idea with the proxy by performing forecasts over four e-folding times. Specifically, we compute forecast errors for five independent forecasts and average the results, shown in Table 2.4. The ranking of (average) long-range forecast errors generally reflects that of the short-range forecasts (see Table 2.3), but with OmFs now clearly distinguishing between

Table 2.4: Long-range forecast error (four e-folding times). Listed are average forecast errors (OmFs and TmFs), scaled by the respective macroscopic errors. The large ensemble EnKF uses $N_e = 1000$, all localized/inflated EnKFs use $N_e = 100$.

	Large Ens.	Shrinkage	Climatological	Wave-vector	Wave-number	SEC
OmF	0.17	0.29	0.41	0.83	0.74	0.69
TmF_A	0.18	0.30	0.43	0.83	0.75	0.69
TmF_ω	0.18	0.29	0.48	0.80	0.93	0.65

the better performing localizations. We can now identify the shrinkage scheme as the most appropriate localization scheme, based only on observed quantities (OmFs) – we no longer rely on TmFs for that conclusion. In other words, our earlier experiments, with short range forecasts, suggest that shrinkage and climatological localization perform similarly in terms of errors in observed quantities, but when taking TmFs into account, the shrinkage scheme has clear advantages. The longer range forecasts reveal the advantages *without* considering TmFs. This suggests that long range forecasts might indeed be an effective tool to identify useful localization schemes in geomagnetic DA.

Shortened spin-up with localized DA

In an effective EnKF system, forecast errors from an uninformed, initial ensemble are typically reduced over several assimiliation cycles before settling around a minimum. This period of transition from large to small forecast errors is referred to as the *spin-up time*. We previously examined average forecast errors only after the spin-up time. This is typical in many applications of DA, e.g., NWP, where one may spin-up only once, because new observations are routinely collected and their assimilation maintains accuracy. In geomagnetic DA, fundamental aspects of, e.g., the assimilation schemes, model parameters, and observational records being used, are often the focus of new studies (see, e.g., Tangborn & Kuang 2015; Sanchez et al. 2020). For this reason, assimilations frequently begin with relatively uninformative priors and must undergo a spin-up period. Given the limited observational record of the geomagnetic field, the rate of

error reduction through initial assimilations is relevant and a short spin-up time is essential: If the spin-up time is long, the observational record may be insufficient for achieving the sought-after error levels and/or carrying out desired validation experiments. We now consider forecast errors during the spin-up time, for the purpose of determining how rapidly forecast errors are reduced from an initial ensemble.

Specifically, we compare the spin-up times of EnKFs with and without localization/ inflation, using a large ensemble size ($N_e = 1000$) for the unlocalized EnKF and $N_e = 100$ for the localized EnKFs. The TmFs in the magnetic and velocity fields over the first 200 assimilation cycles are shown in Figure 2.8. The average rate of decrease in forecast error over the initial assimilations varies among the runs with the large ensemble (blue) clearly, and unsurprisingly, being the most rapid. Generally, the shrinkage scheme (green) produces the lowest, or among the lowest errors, during spin-up time. The climatological scheme (orange) decreases slowly but steadily while the SEC (yellow) initially struggles along with the all-around, poor performing wave-number scheme (red).

For a more objective analysis, we record the number of assimilation cycles required to reduce forecast error to 1% of the macroscopic error in each of the two fields. The results are shown in Table 2.5 and serve to summarize and synthesize the time series in Figure 2.8. We find that an EnKF with a large ensemble requires 14 assimilation cycles to reduce error to 1% in the magnetic field, and 25 to similarly reduce errors in the velocity. The localized EnKFs require more assimilations to achieve similarly small errors. The shortest spin-up with $N_e = 100$ is achieved by an EnKF with shrinkage localization, followed by wave-vector localization, SEC, and climatological localization. The wave-number scheme takes the longest to reduce error to 1% in the magnetic field and never really stabilizes in the velocity field. We thus see that the localization schemes that lead to the smallest errors also reduce the overall spin-up time of an EnKF. Localization and inflation are thus not only useful to bring down the "stationary" errors, but they are also critical to reduce the number of assimilation cycles one has to do before one

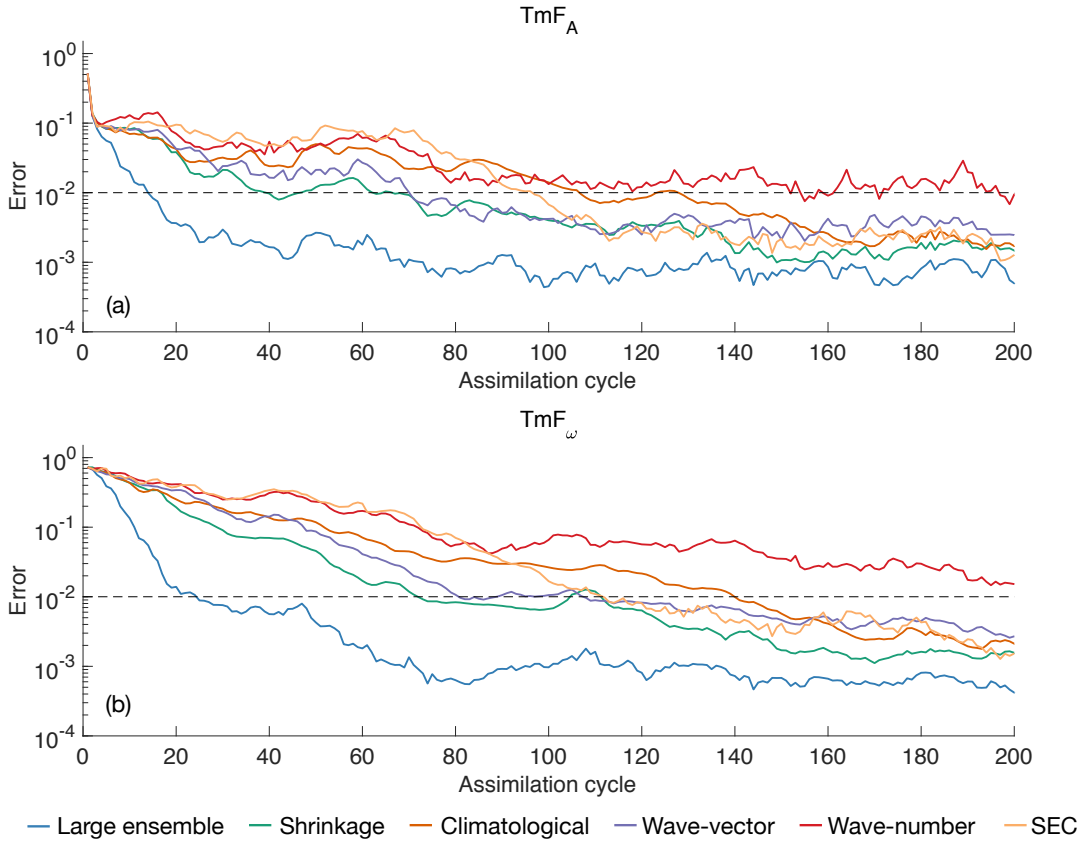

Figure 2.8: Forecast errors (TmF) as a function of the first 200 assimilation cycles for various EnKFs with and without localization/inflation. (a) Error in magnetic field. (b) Error in the velocity field. In both panels the blue line results from a run using a large ensemble of $N_e = 1000$ with no localization or inflation. The localized and inflated EnKFs (scheme indicated by color) use a smaller ensemble size of $N_e = 100$. All errors are scaled by the macroscopic errors. The black dashed line indicates an error level of 1% of the macroscopic error.

Table 2.5: Assimilation cycles required to reduce forecast errors in the magnetic field (first row) and velocity field (second row) to 1% of the macroscopic error. The large ensemble EnKF uses an ensemble size $N_e = 1000$ and no localization or inflation. The remaining EnKFs use localization and inflation (scheme indicated) and an ensemble size $N_e = 100$.

	Large Ens.	Shrinkage	Climatological	Wave-vector	Wave-number	SEC
Magnetic field	14	40	107	71	152	97
Velocity	25	72	140	81	-	112

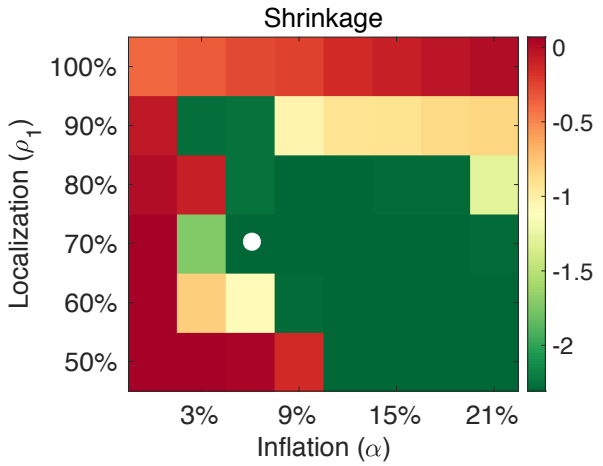

Figure 2.9: Base 10 logarithm of OmFs, scaled by macroscopic error, of an EnKF with ensemble size $N_e = 100$ and different localization (ρ) and inflation (α) parameters (shrinkage localization). Red indicates ineffective DA (macroscopic error); green indicates an effective DA with OmFs comparable to those of an EnKF with a much larger ensemble. A white dot indicates the parameters chosen for further experiments.

reaches acceptable error levels.

Localization vs. inflation and details of the tuning process

We discuss some details of the tuning of localization and inflation, and how to determine an "optimal" localization scheme by resolving a trade-off between localization and inflation. First, a typical gridded error, as obtained during the tuning of localization and inflation, is shown in Figure 2.9. Red regions contain localization and inflation parameters that result in macroscopic errors (OmF ≈ 1); green region contains the localization and inflation parameters that result in small OmF (OmF $\approx 10^{-3}$–10^{-2}). We notice a pattern that less aggressive localization (smaller ρ)

leads to more aggressive inflation (larger α). For example, at $\rho = 90\%$, inflation is $\alpha = 3\%$, but at $\rho_1 = 50\%$ the inflation is $\alpha = 12\%$. We observe this pattern in all four schemes we tried. The fifth scheme, the SEC, is adaptive and does not have a tunable localization parameter. We thus only tuned inflation and found that only a large inflation could reduce the OmF. This is indeed in line with what we find with the non-adaptive localization schemes: The SEC is not a very "aggressive" localization scheme and, for that reason, requires relatively large inflation.

The pattern, or trade-off, between localization and inflation is in fact an important detail: Larger values of inflation mean that less weight is given to the prior in each assimilation and that the ensemble is pushed more towards the observations than with lower inflation. Thus, small OmFs in the presence of increasingly large inflation can simply be the result of pulling the observed part of a model closer to the observations, while estimates of unobserved quantities become poorer (over-fitting). For this reason, in an operational assimilation system where one only has access to OmFs as a measure of forecast accuracy, it is desirable to minimize the level of inflation used. Following this line of thinking, we base our notion of optimal localization and inflation on successfully reducing OmF at low inflation.

We further observed that localization and inflation combinations near regions of large OmF may not be useful. For example, in Figure 2.9, values of $\rho_1 = 90\%, \alpha = 3\%$ are successful in reducing OmFs, but when applied to a different nature run this combination failed to reduce OmF. To avoid localization and inflation which may be unreliable in this way, we define optimality by the localization with lowest inflation, that successfully reduces OmF, while not being in the immediate vicinity of macroscopic error. For example, from Figure 2.9 we define the combination of values indicated by the white dot at $\rho_1 = 70\%, \alpha = 6\%$ to be "optimal."

Summary and conclusions drawn from the proxy on the square

We summarize the main conclusions supported by the numerical experiments with the proxy on the square.

(i) The EnKF can drastically reduce forecast errors (by two orders of magnitude).

(ii) Localization and inflation dramatically reduce the computational cost of the EnKF because a small ensemble size is sufficient (in the proxy localization reduced the ensemble size from $N_e \approx 800$ to $N_e \approx 100$).

(iii) Localization and inflation reduce the spin-up time of an EnKF (the number of assimilation cycles required to reach a low error level).

(iv) The details of the localization are marginally important when one considers observed quantities (OmF) over short forecast horizons, but are critical when one also takes unobserved quantities (TmF) into consideration.

(v) Long range forecasts can be used to clarify differences between localization schemes that lead to similar errors in observed quantities, but very different errors in unobserved quantities. More specifically, longer-range forecast are an effective tool to identify useful localization schemes based on observed quantities only.

The reasons for (iii) and (iv) are that during localization, one assumes an underlying correlation structure, and this assumed structure may be appropriate or not. We found that four out of five of the localization schemes we tried lead to a clear reduction in errors, but none of the schemes leads to errors that are comparable to those of a large ensemble EnKF. Thus, the schemes do not capture all relevant aspects of the underlying correlation structure. Many localization schemes may indeed exhibit similar performance when considering errors in observed quantities; longer range forecasts can be used to identify appropriate localization schemes.

Overall, we found that shrinkage localization is most effective. This is perhaps surprising because shrinkage localization draws on no motivation from the physics or statistics of the proxy model, but rather broadly suppresses correlations at a global level. Indeed, we find that when we rank the localization schemes by increasing errors, the schemes dampen fewer correlations. For example, the climatological localization suppresses correlations except for a few between

certain modes of A and ω with the same wave-number vector. Thus, climatological localization performs similarly to shrinkage localization, because both schemes significantly suppress the vast majority of correlations. Wave-vector localization largely allows correlations between modes with similar wave-number vectors, whether they represent A or ω. Thus, the correlations suppressed by wave-vector localization are a subset of those suppressed by the climatological or shrinkage schemes, and we observe that forecast errors are slighty larger for wave-vector localization than for shrinkage or climatological localization. Wave-number localization allows correlations between modes of similar total wave number, and, therefore, permits correlations similar to those of the wave-vector scheme *and* correlations between modes of similar length scales, but different orientations. For this reason, wave-number localization suppresses the least amount of correlation and indeed exhibits the largest errors. Thus, the success of a localization technique is, at least in part, related to the extent to which it provides a blanket suppression of all correlations and, as a consequence, limits the damage of spurious correlations.

To gain further insights, we consider the EnKF with a large ensemble size ($N_e = 1000$, no localization or inflation), and compute the ensemble correlation between an observed mode of A with wave-vector $\mathbf{k} = \hat{\mathbf{x}}$ and an unobserved, high-energy mode of ω, with wave vector $\mathbf{k} = 2\hat{\mathbf{x}} - 4\hat{\mathbf{y}}$. We take this large ensemble as an indicator of what the relevant correlation structure may be. The time series of correlation, along with a histogram of the correlation, is shown in Figure 2.10. We note that while the wave vectors are certainly of different direction and magnitude, the large ensemble regularly indicates strong correlations between these modes. Moreover, the climatological correlation of these two modes is -0.02, which differs quite a bit from the actual correlation at many instances. At time instances when the climatological correlation differs from the instantaneous correlation, we expect that climatological localization is ineffective because it dampens a relevant correlation. Indeed, we observe this behavior not only in the two modes shown here. The large ensemble indicates strong correlations between many modes which are dissimilar in wave-vector/number and show no climatological correlation. This occurs because

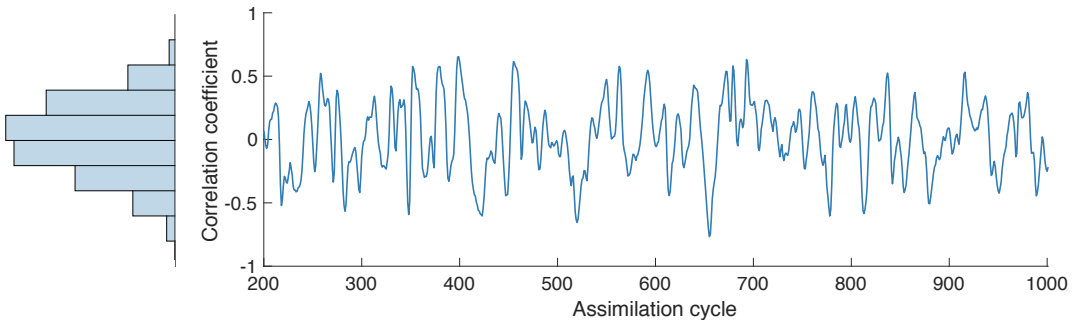

Figure 2.10: Forecast ensemble correlation between an observed mode of A and an unobserved mode of ω for 800 forecasts during a DA run using an ensemble of $N_e = 1000$ and no localization or inflation. The left portion of the figure provides a histogram illustrating the distribution of the correlations. Despite the significant correlations seen above, the modes have shown no strong climatological correlation and differ in wave-number/vector.

over short-term forecasts, perturbations propagate through the spectrum in complicated ways which can not necessarily be captured by the similarity of modes, or the long-term correlations within the system. For example, consider the Fourier expansions of A and ω

$$A = \sum_{k_x,k_y} \hat{A}_{k_x,k_y} e^{2\pi i(k_x+k_y)}, \quad \omega = \sum_{k_x,k_y} \hat{\omega}_{k_x,k_y} e^{2\pi i(k_x+k_y)}, \tag{2.41}$$

where $\hat{A}_{k_x,k_y}$ and $\hat{\omega}_{k_x,k_y}$ are the respective Fourier coefficients. Substituting into equations (2.17)-(2.19) it can be seen that the evolution of the Fourier coefficient $\hat{A}_{1,0}$ is described by

$$\frac{d\hat{A}_{1,0}}{dt} = \sum_{k_x^2+k_y^2\neq 0} \left[\frac{k_y}{k_x^2+k_y^2} \hat{\omega}_{k_x,k_y} \hat{A}_{1-k_x,-k_y} \right] - \frac{4\pi^2}{R_m} \hat{A}_{1,0} + \frac{\hat{\omega}_{1,0}}{2\pi} i. \tag{2.42}$$

The last two terms of the right hand side of (2.42) come from the linear terms of (2.19). We note that in the climatological covariance, the real and imaginary parts of $\hat{A}_{1,0}$ only exhibit strong correlations with the imaginary and real parts of $\hat{\omega}_{1,0}$, respectively. This is a natural consequence of the last term being proportional to $i\hat{\omega}_{1,0}$. The summation is the result of the non-linear term $-(\mathbf{v}\cdot\nabla)A$, and it is here that we see the wide-ranging mixture of modes which can influence $\hat{A}_{1,0}$. While the long-time variations in two modes may be uncorrelated, it is clear

from (2.42) that perturbations in, e.g., $\hat{\omega}_{2,-4}$ can propagate to $\hat{A}_{1,0}$ over short-term forecasts (see figure 2.10). Moreover, the influence of the perturbation on $\hat{A}_{1,0}$ is dependent not only on the sign and magnitude of $\hat{\omega}_{2,-4}$, but also that of $\hat{A}_{-1,4}$. The result is that the structure of the forecast covariances vary with the states of the ensemble members. Indeed, since the correlation structure seems to vary, adaptive localization schemes might indeed be a constructive way forward (with hopes that adaptive strategies can be devised that capture the relevant correlations, which is *not* an easy task, as indicated by the relatively poor performance of the adaptive SEC scheme).

The above findings and discussion are immediately applicable to geomagnetic DA, where, thus far, only a variant of climatological localization has been applied to an ensemble-based system (Sanchez et al. 2019, 2020). Our results suggest that it may be well worthwhile to test a simple shrinkage scheme and, if one uses OSSEs (not actual geomagnetic data), to check errors in unobserved components of the geodynamo. In an operational scheme, where one is restricted to observed quantities, experiments with longer range forecasts may turn out to be important.

2.4.3 Results: proxy on the sphere

The numerical experiments with the proxy on the square suggest that the shrinkage and climatological localization schemes are most promising and we now test these on the spherical proxy. We tune localization and inflation as before and consider an ensemble of size $N_e = 100$. The nature run is 301 DA cycles long and we compute average errors (OmF and TmF) over the last 50 cycles to account for spin-up. Table 2.6 lists the errors (OmF and TmF) of the tuned EnKFs. We note that we tried EnKF without localization or inflation with ensemble size $N_e = 1000$, but this configuration still leads to macroscopic error. This means that the required ensemble size for reliable DA *without* localization or inflation is larger than 1000 (but our experiments are not sufficient for determining what the minimum required ensemble size really is, we only found a lower bound). The fact that the localized and inflated EnKFs lead to drastically reduced errors thus, again, speaks for the relevance of localization and inflation to keep the computational budget

Table 2.6: Average errors (OmF and TmFs), scaled by the respective macroscopic errors and multiplied by 10^{-3} (proxy on the sphere). Both EnKFs are run with ensemble size $N_e = 100$.

		Magnetic field			Velocity	
	OmF	$\text{Spread}^f_{\text{obs}}$	TmF^f_A	Spread^f_A	TmF^f_ω	Spread^f_ω
Shrinkage	4.9	5.6	2.1	3.2	3.3	9.5
Climatological	5.5	6.3	3.2	4.9	8.6	18.4

in check.

As observed in the experiments with the proxy on the square, we find that both localization schemes are effective and reduce the errors by more than two orders of magnitude and that ensemble spread is comparable to the errors (suggesting a well-tuned DA system). Overall, we find the smallest errors in the shrinkage scheme, which again is consistent with what we found for the proxy on the square. We note, however, that the errors in the unobserved velocity field is much larger for the climatological localization than for the shrinkage method. On the spherical proxy, shrinkage localization seems to have a clear-cut advantage over climatological localization. This provides further support to the notion that there exists a structure to correlations between modes, in short-term forecasts of the model, which is more complicated than indicated by the climatological statistics.

The advantages of shrinkage localization can further be highlighted by considering longer-range forecasts (see experiments with the proxy on the square). We initialize forecasts of four e-folding times with the ensemble mean of the two Kalman filters (with climatological and shrinkage localization). The results are illustrated in Figure 2.11. We see that the forecast error of the shrinkage scheme is much smaller than the forecast error of climatological localization. The reason is that climatological localization leads to larger errors in the unobserved state (the velocity), which are amplified relatively quickly by the dynamics. Indeed, on longer range forecasts, we note a clear reduction in forecast errors in observed *and* unobserved quantities. This corroborates the idea that longer range forecast errors can be used as a tool to diagnose the applicability of localization methods: An appropriate localization scheme should lead to smaller

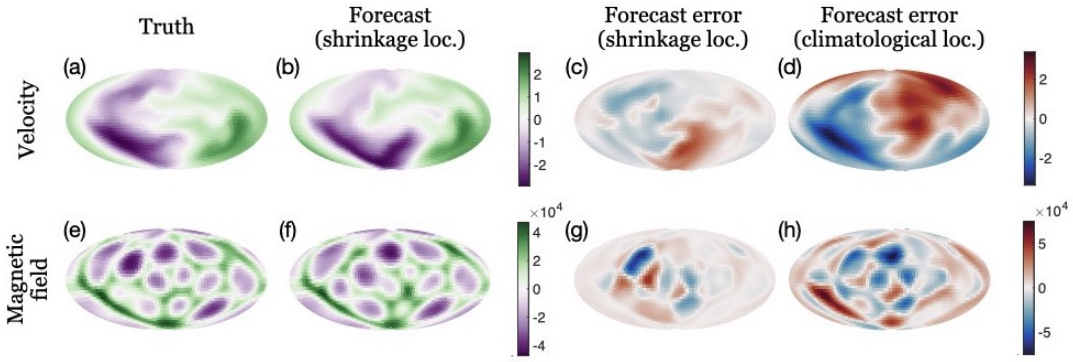

Figure 2.11: A comparison of forecast skill when using initial conditions estimated using the shrinkage and climatological localization schemes. (a) true state of A, (b) forecast of A from shrinkage localization (c) error in forecast of A from shrinkage localization, (d) error in forecast of A from climatological localization, (e) true state of ω, (f) forecast of ω from shrinkage localization (g) error in forecast of ω from shrinkage localization, (h) error in forecast of ω from climatological localization.

Table 2.7: Assimilation cycles required to reduce forecast errors in the magnetic field (first row) and velocity field (second row) to 1% of the macroscopic error using shrinkage (second column) and climatological localization (third column).

	Shrinkage	Climatological
Magnetic field	137	170
Velocity	179	212

longer-range forecast errors.

Finally, we compare the spin-up times of the EnKFs using the shrinkage and climatological localization schemes on the sphere. Beginning with an uniformed, initial ensemble, we record the number of assimilations needed to reduce the error to 1% of the macroscopic error in each field. The results are shown in Table 2.7 where, as observed with the proxy on the square, the shrinkage scheme requires fewer assimilations to reduced errors. Thus, we again find that the choice of localization scheme not only determines the average forecast errors over a long DA run, it also impacts the the number of assimilations necessary to reduce initial errors to a desired level.

2.5 Summary and conclusions

We designed a simplified proxy model to represent some of the numerical and computational challenges in geomagnetic DA. The proxy swaps the Navier-Stokes equation for a simpler, but chaotic flow (the Kuramoto-Sivashinsky equation), and couples the flow variables to a magnetic field via an induction equation and a Lorentz-like coupling term. The proxy model is computationally simple enough to enable systematic numerical studies and data assimilation experiments.

The strategy of using proxy models to rapidly prototype and test numerical DA schemes has proven useful and effective in NWP. Specifically, we designed the proxy to replicate the following challenges in geomagnetic DA:

(i) The proxy is a chaotic dynamic system, consisting of two coupled PDEs for magnetic and velocity fields;

(ii) the proxy is amenable to spectral discretization;

(iii) magnetic and velocity fields are coupled, but only the spectrum of the magnetic field is partially observed;

(iv) the proxy model is dynamically rich enough to require localization and inflation for efficient numerical DA.

Due to the simplifications, results obtained with the proxy model are qualitative, not quantitative and one must carefully evaluate how results obtained with the proxy model may extend to realistic, or real, geomagnetic DA systems.

The proxy model has a 2D geometry to keep computations simple and we present versions of the proxy on square and spherical geometries. The proxy on the square makes use of FFT and, for that reason, is computationally simpler than the proxy on the sphere which requires spherical harmonic transforms. This suggests using the proxies hierarchically: The proxy on the square can be use to screen for possibly useful techniques, and one can then perform a selected

set of experiments on the spherical proxy. Using the model hierarchy, we ran a large number of systematic DA experiments with the goal to investigate the role of localization and inflation within geomagnetic DA. The numerical experiments suggest the following key findings.

(a) Localization and inflation are invaluable for efficient geomagnetic DA because localized DA can keep the errors small, even if the ensemble size is also small;

(b) localization and inflation can reduce the spin-up period, i.e., the number of DA cycles required to reach a steady and low error level;

(c) a simple shrinkage localization scheme may be more appropriate than popular climatological localization methods, because the correlation structure relevant for DA may be very different from the climatological correlations;

(d) many localization schemes may produce similar errors in observed quantities on short time scales, but errors in the unobserved quantities can be quite different for the different schemes;

(e) performing longer range forecasts may prove to be a useful tool for determining how appropriate a localization scheme really is.

A natural next step for continuing this work, and to determine the actual usefulness of the proxy, is to test some of our hypothesis in a more realistic or real geomagnetic DA system.

DATA AVAILABILITY

The code for the model on the square (implemented in Matlab) is available at https://github.com/kjg136/GDA_Testbed (curated version). The code for the model on the sphere cannot be made publicly available at this time as it relies on a spherical harmonic transform owned by a third party and used with their permission. All numerical results, and the code used for the model on the square in this paper, are at https://zenodo.org/record/5304367#.YSpZwtNKhTY.

Acknowledgements

KG acknowledges that this work was supported by NASA Headquarters under the NASA Earth and Space Science Fellowship Program - Grant "80NSSC18K1351". MM and KG are supported by the US Office of Naval Research (ONR) grant N00014-21-1-2309. AT and WK are supported by NASA Earth Surface and Interior Program. We thank Sabrina Sanchez (Max Planck Institute for Solar System Research) and an anonymous reviewer for helping us improve this paper.

All authors contributed to the ideas behind the approach taken in the manuscript with KG taking the lead; all authors contributed to writing the paper, with KG taking the lead and producing the first draft; KG wrote the code and designed and performed the numerical experiments.

Chapter 2, in full, is a reprint of material as it appears in Gwirtz, K., Morzfeld, M., Kuang, W. and Tangborn, A., A testbed for geomagnetic data assimilation, *Geophysical Journal International*, **227** (3), 2180-2203, (2021). The dissertation author was the primary investigator and author of this paper.

References

Ades, M. & van Leeuwen, P. J., 2015. The effect of the equivalent-weights particle filter on dynamical balance in a primitive equation model, *Mon. Weather Rev.*, **143**(2), 581–596.

Anderson, J. L., 2012. Localization and sampling error correction in ensemble Kalman filter data assimilation, *Mon. Weather Rev.*, **140**, 2359–2371.

Anderson, J. L., 2016. Reducing correlation sampling error in ensemble Kalman filter data assimilation, *Mon. Weather Rev.*, **144**, 913–925.

Anderson, J. L. & Anderson, S. L., 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, *Mon. Weather Rev.*, **127**(12), 2741 – 2758.

Armbruster, D., Guckenheimer, J., & Holmes, P., 1989. Kuramoto-Sivashinsky dynamics on the center-unstable manifold, *SIAM J Appl Math.*, **49**(3), 676–691.

Bärenzung, J., Wicht, M. H. J., Lesur, V., & Sanchez, S., 2020. The Kalmag model as a candidate for IGRF-13, *Earth, Planets and Space*, **72**(163).

Barrois, O., Hammer, M. D., Finlay, C. C., Martin, Y., & Gillet, N., 2018. Assimilation of ground and satellite magnetic measurements: inference of core surface magnetic and velocity field changes, *Geophys. J. Int.*, **215**(1), 695–712.

Bauer, P., Thorpe, A., & Brunet, G., 2015. The quiet revolution of numerical weather prediction, *Nature*, **252**, 45–55.

Bonavita, M., Isaksen, L., & Hólm, E., 2012. On the use of EDA background error variances in the ECMWF 4D-Var, *Q. J. R. Meteorol. Soc.*., **138**(667), 1540–1559.

Browne, P. A., 2016. A comparison of the equivalent weights particle filter and the local ensemble transform Kalman filter in application to the barotropic vorticity equation, *Tellus A*, **68**, 30466.

Buehner, M., McTaggart-Cowan, R., & Heilliette, S., 2017. An ensemble Kalman filter for numerical weather prediction based on variational data assimilation: VarEnKF, *Mon. Weather Rev.*, **145**(2), 617 – 635.

Chorin, A. J. & Morzfeld, M., 2013. Conditions for successful data assimilation, *J. Geophys. Res. Atmos.*, **118**(20), 11,522–11,533.

Chorin, A. J., Morzfeld, M., & Tu, X., 2013. Implicit sampling, with application to data assimilation, *Chin. Ann. Math. Ser. B*, **34**, 89–98.

Courtier, P., 1997. Variational methods, *J. Meteorol. Soc. Japan*, **75**(1B), 211–218.

Cowling, T. G., 1933. The Magnetic Field of Sunspots, *Mon. Not. R. Astron. Soc.*, **94**(1), 39–48.

Cox, S. & Matthews, P., 2002. Exponential time differencing for stiff systems, *J. Comput. Phys.*, **176**, 430–455.

Errico, R. M., Yang, R., Priv, N. C., Tai, K.-S., Todling, R., Sienkiewicz, M. E., & Guo, J., 2013. Development and validation of observing-system simulation experiments at NASA's global modeling and assimilation office, *Q. J. R. Meteorol. Soc.*., **139**(674), 1162–1178.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res. Oceans*, **99**(10), 10143–10162.

Evensen, G., 2006. *Data assimilation: the ensemble Kalman filter*, Springer.

Fornberg, B., 1996. *A Practical Guide to Pseudospectral Methods*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press.

Fournier, A., Eymin, C., & Alboussière, T., 2007. A case for variational geomagnetic data assimilation: insights from a one-dimensional, nonlinear, and sparsely observed MHD system, *Nonlinear Process. Geophys.*, **14**(2), 163–180.

Fournier, A., Nerger, L., & Aubert, J., 2013. An ensemble Kalman filter for the time-dependent analysis of the geomagnetic field, *Geochem. Geophys. Geosyst.*, **14**, 4035–4043.

Gharamti, M. E., Raeder, K., Anderson, J., & Wang, X., 2019. Comparing adaptive prior and posterior inflation for ensemble filters using an atmospheric general circulation model, *Mon. Weather Rev.*, **147**(7), 2535 – 2553.

Hamill, T. M., Whitaker, J. S., Anderson, J. L., & Snyder, C., 2009. Comments on "Sigma-Point Kalman Filter Data Assimilation Methods for Strongly Nonlinear Systems", *J. Atmos. Sci.*, **66**(11), 3498–3500.

Harty, T., Morzfeld, M., & Snyder, C., 2021. Eigenvector-spatial localisation, *Tellus A*, **73**(1), 1–18.

Hoffman, R. N. & Atlas, R., 2016. Future observing system simulation experiments, *Bull. Am. Meteorol. Soc.*, **97**(9), 1601 – 1616.

Holm, H. H., Sætra, M. L., & van Leeuwen, P. J., 2020. Massively parallel implicit equal-weights particle filter for ocean drift trajectory forecasting, *J. Comput. Phys.*, **6**, 100053.

Hooper, A. P. & Grimshaw, R., 1985. Nonlinear instability at the interface between two viscous fluids, *Phys. Fluids*, **28**(1), 37–45.

Hulot, G., Lhuillier, F., & Aubert, J., 2010. Earth's dynamo limit of predictability, *Geophys. Res. Lett.*, **37**, L06305.

Hunt, B. R., Kostelich, E. J., & Szunyogh, I., 2007. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, *Physica D*, **230**(1), 112–126, Data Assimilation.

Jardak, M., Navon, I. M., & Zupanski, M., 2010. Comparison of sequential data assimilation methods for the Kuramoto-Sivashinsky equation, *Int. J. Numer. Methods Fluids*, **62**(4), 374–402.

Kalogirou, A., Keaveny, E. E., & Papageorgiou, D. T., 2015. An in-depth numerical study of the two-dimensional Kuramoto-Sivashinsky equation, *Proc. R. Soc. A.*, **471**.

Kotsuki, S., Ota, Y., & Miyoshi, T., 2017. Adaptive covariance relaxation methods for ensemble data assimilation: experiments in the real atmosphere, *Q. J. R. Meteorol. Soc.*., **143**(705), 2001–2015.

Kuang, W. & Bloxham, J., 1999. Numerical modeling of magnetohydrodynamic convection in a rapidly rotating spherical shell: Weak and strong field dynamo action, *J. Comput. Phys.*, **153**(1), 51–81.

Kuramoto, Y. & Tsuzuki, T., 1975. On the formation of dissipative structures in reaction-diffusion systems, *Prog. Theor. Phys.*, **54**(3), 687–699.

Langel, R. A. & Estes, R. H., 1982. A geomagnetic field spectrum, *Geophys. Res. Lett.*, **9**(4), 250–253.

Li, H., Kalnay, E., & Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter, *Q. J. R. Meteorol. Soc.*., **135**(639), 523–533.

Li, K., Jackson, A., & Livermore, P. W., 2011. Variational data assimilation for the initial-value dynamo problem, *Phys. Rev. E*, **84**, 056321.

Li, K., Jackson, A., & Livermore, P. W., 2014. Variational data assimilation for a forced, inertia-free magnetohydrodynamic dynamo model, *Geophys. J. Int.*, **199**, 1662–1676.

Lorenz, E. N., 1963. Deterministic nonperiodic flow, *J. Atmos. Sci.*, **20**(2), 130 – 141.

Lorenz, E. N., 1995. Predictability: a problem partly solved., *ECMWF Seminar Proceedings on Predictability*, **1**, 118.

Lunderman, S., Morzfeld, M., & Posselt, D. J., 2021. Using global Bayesian optimization in ensemble data assimilation: parameter estimation, tuning localization and inflation, or all of the above, *Tellus A*.

Minami, T., Nakano, S., Lesur, V., Takahashi, F., Matsushima, M., Shimizu, H., Nakashima, R., Taniguchi, H., & Toh, H., 2020. A candidate secular variation model for IGRF-13 based on MHD dynamo simulation and 4DEnVar data assimilation, *Earth, Planets Space*, **72**(136).

Morzfeld, M., Hodyss, D., & Snyder, C., 2017. What the collapse of the ensemble Kalman filter tells us about particle filters, *Tellus A*, **69**(1), 1283809.

Morzfeld, M., Adams, J., Lunderman, S., & Orozco, R., 2018. Feature-based data assimilation in geophysics, *Nonlinear Process. Geophys.*, **25**(2), 355–374.

Papageorgiou, D., Maldarelli, C., & Rumschitzki, D., 1990. Nonlinear interfacial stability of core-annular film flows, *Phys. Fluids A*, **2**(3), 340–352.

Ropp, G., Lesur, V., Bärenzung, J., & Holschneider, M., 2020. Sequential modelling of the Earth's core magnetic field, *Earth, Planets and Space*, **72**(153).

Sabaka, T. J., Tøffner-Clausen, L., Olsen, N., & Finlay, C. C., 2020. CM6: a comprehensive geomagnetic field model derived from both CHAMP and Swarm satellite observations, *Earth, Planets and Space*, **72**(80).

Sanchez, S., Fournier, A., Aubert, J., Cosme, E., & Gallet, Y., 2016. Modelling the archaeomagnetic field under spatial constraints from dynamo simulations: a resolution analysis, *Geophys. J. Int.*, **207**(2), 983–1002.

Sanchez, S., Wicht, J., Bärenzung, J., & Holschneider, M., 2019. Sequential assimilation of geomagnetic observations: perspectives for the reconstruction and prediction of core dynamics, *Geophys. J. Int.*, **217**(2), 1434–1450.

Sanchez, S., Wicht, J., & Bärenzung, J., 2020. Predictions of the geomagnetic secular variation based on the ensemble sequential assimilation of geomagnetic field models by dynamo simulations, *Earth Planets Space*, **72**(157).

Shlyaeva, A., Whitaker, J., & Snyder, C., 2019. Model-space localization in serial ensemble filters, *J. Adv. Model. Earth Syst.*, **11**(6), 1627–1636.

Sivashinsky, G., 1977. Nonlinear analysis of hydrodynamic instability in laminar flames-I. derivation of basic equations, *Acta Astronaut.*, **4**, 1177–1206.

Sun, Z. & Kuang, W., 2015. An ensemble algorithm based component for geomagnetic data assimilation, *Terr. Atmospheric Ocean. Sci.*, **26**.

Sun, Z., Tangborn, A., & Kuang, W., 2007. Data assimilation in a sparsely observed one-dimensional modeled MHD system, *Nonlinear Process. Geophys.*, **14**, 181–192.

Tangborn, A. & Kuang, W., 2015. Geodynamo model and error parameter estimation using geomagnetic data assimilation, *Geophys. J. Int.*, **200**(1), 664–675.

Tangborn, A. & Kuang, W., 2018. Impact of archeomagnetic field model data on modern era geomagnetic forecasts, *Phys. Earth Planet. Inter.*, **276**, 2 – 9, Special Issue:15th SEDI conference.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., & Whitaker", J. S., 2003. Ensemble square root filters, *Mon. Weather Rev.*, **131**, 1485–1490.

Touloumis, A., 2015. Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings, *Computational Statistics & Data Analysis*, **83**, 251–261.

Zeng, X., Atlas, R., Birk, R. J., Carr, F. H., Carrier, M. J., Cucurull, L., Hooke, W. H., Kalnay, E., Murtugudde, R., Posselt, D. J., Russell, J. L., Tyndall, D. P., Weller, R. A., & Zhang, F., 2020. Use of observing system simulation experiments in the United States, *Bull. Am. Meteorol. Soc.*, **101**(8), E1427 – E1438.

Zhang, M. & Zhang, F., 2012. E4DVar: Coupling an ensemble Kalman filter with four-dimensional variational data assimilation in a limited-area weather prediction model, *Mon. Weather Rev.*, **140**(2), 587 – 600.

Zhen, Y. & Zhang, F., 2014. A probabilistic approach to adaptive covariance localization for serial ensemble square root filters, *Mon. Weather Rev.*, **142**(12), 4499 – 4518.

Chapter 3

Can one use Earth's magnetic axial dipole field intensity to predict reversals?

K. Gwirtz*, M. Morzfeld*, A. Fournier^o, G. Hulot^o
* Cecil H. and Ida M. Green Institute of Geophysics and Planetary Physics,
Scripps Institution of Oceanography, University of California, San Diego;
^{o} Université de Paris, Institut de Physique du Globe de Paris, CNRS, F-75005 Paris, France.
doi: 10.1093/gji/ggaa542

Abstract. We study predictions of reversals of Earth's axial magnetic dipole field that are based solely on the dipole's intensity. The prediction strategy is, roughly, that once the dipole intensity drops below a threshold, then the field will continue to decrease and a reversal (or a major excursion) will occur. We first present a rigorous definition of an intensity threshold-based prediction strategy and then describe a mathematical and numerical framework to investigate its validity and robustness in view of the data being limited. We apply threshold-based predictions to a hierarchy of numerical models, ranging from simple scalar models to 3D geodynamos. We find that the skill of threshold-based predictions varies across the model hierarchy. The differences in skill can be explained by differences in how reversals occur: if the field decreases towards a reversal slowly (in a sense made precise in this paper), the skill is high, and if the field decreases quickly, the skill is low. Such a property could be used as an additional criterion to identify which models qualify as Earth-like. Applying threshold-based predictions to Virtual Axial Dipole Moment (VADM) paleomagnetic reconstructions (PADM2M and Sint-2000) covering the last

two million years, reveals a moderate skill of threshold-based predictions for Earth's dynamo. Besides all of their limitations, threshold-based predictions suggests that no reversal is to be expected within the next 10 kyr. Most importantly, however, we show that considering an intensity threshold for identifying upcoming reversals is intrinsically limited by the dynamic behavior of Earth's magnetic field.

3.1 Introduction

Earth possesses a time-varying magnetic field which is generated and sustained by turbulent flow of liquid metal alloy in the core. The field varies over a wide range of spatial and temporal scales, but this paper focuses on the dynamics of the axial dipole component over millions of years (Myr henceforth), which are relevant for the investigation of dipole reversals. When a reversal occurs, the intensity of the dipole collapses and then builds up in reversed polarity, with the magnetic north pole becoming the south pole and vice versa. Occurrence of dipole reversals is well-documented over the past 150 Myr (Ogg 2012; Cande & Kent 1995; Lowrie & Kent 2004). We thus know that the last reversal occurred about 780 kilo years (kyr henceforth) ago and that the average reversal rate over the past 5-10 Myr is about 4 reversals per Myr (see, e.g., Morzfeld & Buffett 2019). Given these numbers, we wonder if we can reliably predict if a reversal can be expected to occur any time soon.

At first sight, the task seems hopeless because simulations of Earth's magnetic field suggest that the geomagnetic field is not predictable beyond a century (Hulot et al. 2010b; Lhuillier et al. 2011a). The typical time elapsed between two reversals is much larger (often hundreds of millennia) which implies that the exact timing of a reversal cannot be predicted until a reversal is just about to happen (see also Hulot & Le Mouël 1994; Lhuillier et al. 2011b). This predictability limit, however, concerns the field in its full detail, and one may be able to identify macroscopic conditions that occur over long timescales that are largely independent

of the detailed morphology of the field. For the remainder of this paper, we assume that the predictability limit for reversals, viewed as macroscopic features, is larger than the predictability limit of the field's details. This is motivated by the rich low-frequency dynamics of the long-term dipole field (Constable & Johnson 2005), and by the recent study of Morzfeld et al. (2017) that suggested this could possibly be the case.

In fact, many researchers have implicitly relied on this assumption and studied precursors of dipole reversals. Examples include careful investigations of the characteristics of past reversals (see, e.g., Valet & Fournier 2016, for a recent review), studying the field structure during reversals and excursions (Brown et al. 2018), studying the cause of the present fast decrease of the dipole field, which could be a precursor for a reversal, (see, e.g., Hulot et al. 2002; Finlay et al. 2016), and computational modeling (see, e.g., Olson et al. 2009). Besides these efforts, no consensus has been reached as to what a reliable precursor for a reversal is (see, e.g., Constable & Korte 2006; Laj & Kissel 2015). This is caused, at least in part, by the fact that simulations and the paleomagnetic record indicate that the details of dipole reversals vary greatly (see, e.g., Hulot et al. 2010a; Glatzmaier & Coe 2015), even if their directional behavior, as recorded by lava flows, shows some degree of similarity from one reversal to the next (Valet et al. 2012).

Here, we re-visit these issues and specifically test the often suggested possibility that a small value of the dipole's strength could be used as a natural indicator of an upcoming dipole reversal. Our predictions do not distinguish between reversals and major excursions that lead to a near or total collapse of the axial dipole field, but end up with the axial dipole rebuilding with the same polarity (see below for why). In this study, we therefore collectively refer to reversals or such excursions as "low-dipole events" (see Section 3.3 for a precise definition of a low-dipole event).

The idea is as follows. During a low-dipole event, the dipole intensity drops to a very low value. Since the intensity is a continuous function, it must have approached this low intensity level continuously. One may thus ask: can one identify a threshold with the property that if this

threshold is passed, the intensity will continue to decay and a low-dipole event will occur during a specified time interval, called the prediction horizon. The prediction horizon is critical to the usefulness of the prediction strategy. A prediction horizon of several million years, for example, is not useful, because a low-dipole event is likely to occur over these timescales. Similarly, a prediction horizon of a few hundred years is not useful because the low-dipole event may be already in full-swing when we catch it. Given that it takes several kyr for a dipole reversal to take place, a useful prediction horizon should be at least several kyr long.

We can now state the question we want to address more precisely:

Can we identify a threshold that is useful for predicting low-dipole events?

We study this question via a hierarchy of models, ranging from simplified, low-dimensional models, to 3D simulations of Earth's magnetic field. We identify, for each model, a threshold by maximizing a skill score that quantifies the skill of the prediction. When identifying a threshold one should keep in mind that the event "a low-dipole event will occur during the prediction horizon" is rare in comparison to the event "no low-dipole event will occur during the prediction horizon" (at least for useful prediction horizons); this is addressed by using well-established skill scores that are robust to imbalances of the occurrence of one event over another. We carefully discuss the numerical robustness of our approach and also study robustness with respect to the duration of the training data that are used to identify a threshold. We then apply the same methodology to paleomagnetic reconstructions and discuss the geophysical implications of our study.

Overall, we introduce a new prediction strategy, apply it to four models and two paleomagnetic reconstructions (PADM2M and Sint-2000), and test it with a variety of skill scores. This causes us to use a large number of acronyms, most of which are listed in a Table 3.3 in the Appendix.

3.2 Background: model hierarchy and skill scores

We briefly describe the geomagnetic models we use and then outline how to assess prediction strategies via skill scores and receiver operator characteristic (ROC) curves. Readers who are familiar with the models we use or with skill scores and ROC curves may skip this part of the paper.

3.2.1 Numerical modeling of the geomagnetic field

A realistic model for the Earth's magnetic field is a three-dimensional magneto-hydrodynamic (MHD) model. Today's MHD models are realistic representations of Earth's magnetic field over a large range of spatial and temporal scales (Schaeffer et al. 2017; Aubert 2019; Wicht & Sanchez 2019), but the simulation of dipole reversals remains a computational challenge and the number of MHD simulations that exhibit reversals remains limited (Lhuillier et al. 2013; Olson et al. 2013). The reason is that Earth-like, high-resolution simulations of the field are difficult to do, even with today's super-computers. As a result, simulations that exhibit reversals often require that they be pushed away from the Earth-like regime. For example, the Ekman number is a control parameter which expresses the ratio of the rotation time scale to the viscous time scale. Increasing the Ekman number amounts to increasing the kinematic diffusivity of the fluid and thereby the laminar character of the simulated flow. This in turn decreases the required resolution and the time-to-solution. For this reason, many reversing simulations are characterized by an Ekman number that is much larger than the Ekman number of the Earth's dynamo.

An alternative to 3D simulations are low-dimensional models. The terminology is perhaps confusing here because the word "dimensional" does not refer to the spatial dimension, but the number of variables within the model. In this terminology, a 3D model is high-dimensional because it contains a large number of variables that describe the three-dimensional structure of the fluid flow and its interactions with the magnetic field. The 3D model we consider below has more

than three million variables and, hence, its dimension is $O(10^6)$. Low-dimensional models aim to represent selected aspects of the geodynamo – in our case the axial dipole over Myr time scales – with only a small number of variables. The models we consider have one or three variables and, hence, dimension one or three – six orders of magnitude less than the 3D model. Examples of low-dimensional models include scalar stochastic differential equations (SDE) that model the time evolution of the axial dipole as a particle in a double well potential (Hoyng et al. 2001; Schmitt et al. 2001; Buffett et al. 2013, 2014; Buffett & Matsui 2015; Buffett 2015; Meduri & Wicht 2016; Morzfeld & Buffett 2019), scalar SDE's that are inspired by MHD (Pétrélis et al. 2009), and systems of chaotic differential equations that model the interaction of the dipole with the non-dipole (quadrupole) field, coupled and perturbed by a velocity variable (Gissinger 2012).

3.2.2 The model hierarchy

We consider three low-dimensional models and one 3D simulation. We give a concise description of all four models we use and refer to the original works for further information. The 3D model we use is unpublished and, for that reason, we provide more information about the 3D model than the simpler models.

The deterministic G12 model

Following Gissinger (2012), we consider the ordinary differential equations

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \mu Q - VD, \quad \frac{\mathrm{d}D}{\mathrm{d}t} = -\nu D + VQ, \quad \frac{\mathrm{d}V}{\mathrm{d}t} = \Gamma - V + QD, \tag{3.1}$$

where $\mu = 0.119$, $\nu = 0.1$, and $\Gamma = 0.9$. Here, D is the dipole and the variable Q represents the quadrupole or, more generally, the non-dipole field; V is a velocity variable that couples D and Q. A change in the sign of D corresponds to a dipole reversal. We refer to this model as the G12 model. A typical simulation with G12 is shown in Figure 3.1. Here, model time t is scaled

to represent the G12 millennium time scale (1 dimensionless time unit = 4 kyr), see Morzfeld et al. (2017). The simulation is done by discretizing the differential equation by a fourth-order Runge-Kutta scheme (Matlab's ode45).

The stochastic P09 model

Pétrélis et al. (2009) derived a model for dipole reversals by considering the interaction of two modes. Using the symmetry of the equations of magnetohydrodynamics $B \to -B$ in an amplitude equation, and by assuming that the amplitude has a shorter time scale than a phase, a stochastic differential equation (SDE) of the form

$$\mathrm{d}x = f(x)\mathrm{d}t + \sqrt{2q}\,\mathrm{d}W, \tag{3.2}$$

is derived for the phase, x, where $f(x)$ and q are defined below. In Equation (3.2), W is Brownian motion, a stochastic process with the following properties: (i) $W(0) = 0$; (ii) $W(t) - W(t+\Delta t) \sim \mathcal{N}(0, \Delta t)$; and (iii) $W(t)$ is almost surely continuous for all $t \geq 0$, see, e.g., Chorin & Hald (2013). Here and below, $\mathcal{N}(m, \sigma^2)$ denotes a Gaussian random variable with mean m, standard deviation σ, and variance σ^2.

More specifically, the SDE for the phase is defined by

$$f(x) = \alpha_0 + \alpha_1 \sin(2x), \quad \sqrt{2q} = 0.2\sqrt{|\alpha_1|}. \tag{3.3}$$

We use the same parameters as in Pétrélis et al. (2009), $\alpha_1 = -185\,\mathrm{Myr}^{-1}$, $\alpha_0/\alpha_1 = -0.9$. The dipole, D, can be calculated from the phase by $D = R\cos(x + x_0)$. Following Morzfeld et al. (2017), we set $x_0 = 0.3$ and $R = 1.3$ (the latter scales the dipole variable D to have approximately the same time average as the relative paleointensity reported by the reconstruction of Sint-2000 (Valet et al. 2005)). For the reminder of this paper, we refer to this model as the P09 model.

Note that the parameters define the model's time scale. The parameters are chosen so

that the P09 model exhibits reversals and excursions, and so that its reversal rate is comparable to that of Earth's dipole. This is illustrated in Figure 3.1, where a typical simulation result with this model is shown. For a simulation we discretize the differential equation using a forward Euler-Maruyama method (Kloeden & Platen 1999). The time step is 1 kyr.

The double well model

A simple model for reversals of a quantity (not necessarily Earth's dipole field) is a particle in a double well potential. Such a model is defined by an SDE model as in equation (3.2), and with an $f(x)$ that is equal to the negative gradient of a double well potential. Variations of this model for geomagnetic dipole reversals have been considered by many researchers (Hoyng et al. 2001; Schmitt et al. 2001; Buffett et al. 2013, 2014; Buffett & Matsui 2015; Buffett 2015; Meduri & Wicht 2016; Morzfeld & Buffett 2019). The basic idea is that the state, x, of the SDE is within one of the two wells of the double well potential and is pushed around by noise (the Brownian motion $\sqrt{2q}\,\mathrm{d}W$). When the noise builds up towards one side of the well, the state may cross over to the other potential well. One can identify a transition from one well to the other as a reversal of Earth's dipole.

We use a recent version of this model, called the Myr model in Morzfeld & Buffett (2019), for which

$$f(x) = \gamma \frac{x}{\bar{x}} \cdot \begin{cases} (\bar{x} - x), & \text{if } x \geq 0 \\ (x + \bar{x}), & \text{if } x < 0 \end{cases}, \tag{3.4}$$

where $\gamma = 0.1$ kyr^{-1}, $\bar{x} = 5.23 \cdot 10^{22}$ Am^2 and $q = 0.34 \cdot 10^{44}$ A^{2}m^4 kyr^{-1}. These parameters define the model's natural time scale and the values we chose are based on configuration (a) in Morzfeld & Buffett (2019), which implies that the model's reversal rate is comparable with Earth's reversal rate. For the reminder of this paper, we refer to this model as the DW model. A typical simulation of this model is shown in Figure 3.1. For a simulation we discretize the equation using a fourth-order Runge-Kutta for the deterministic part and a forward Euler-Maruyama for the

stochastic component (Kloeden & Platen 1999). The time step is 1 kyr.

The 3D model

We consider a three-dimensional, convection-driven, dynamo simulation which exhibits polarity reversals and dipole excursions. The simulation we consider has not yet been published and 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 de Paris, Institut de Physique du Globe de Paris). We refer to this simulation simply as the 3D model for the rest of this paper.

The 3D model uses a pseudo-spectral approximation to solve the set of equations governing rotating dynamo action in a spherical shell geometry (see, e.g., Christensen & Wicht 2015, for details). The scales chosen to non-dimensionalize the set of equations are the same as those used by e.g. Schaeffer et al. (2017). The radius ratio of the inner-core boundary to the core-mantle boundary is set to its present-day value. The 3D model has no-slip boundary conditions on the inner-core and core-mantle boundaries, and it assumes that the inner core is conducting. The four non-dimensional control parameters, as defined e.g. in Schaeffer et al. (2017), are as follows. The Ekman number is 10^{-4}, the Prandtl number is 1, the magnetic Prandtl number is 3 and the Rayleigh number is 15000. These choices result in an average hydrodynamic Reynolds number of 216 (recall that the hydrodynamic Reynolds number is defined as the product of the root-mean-squared velocity by the shell thickness divided by the kinematic viscosity). The open-source, freely available xshells code [1] is used to numerically solve the equations. This code combines the finite difference method in the radial direction with a spherical harmonic representation of field variables in the horizontal direction, using the dedicated SHTns library (Schaeffer 2013). To ensure numerical convergence, a hyperdiffusivity is applied beyond spherical harmonic degree 55.

[1] https://nschaeff.bitbucket.io/xshells/

The resolution of the 3D model is defined by the triplet $(N_r, \ell_{\max}, m_{\max})$, giving the number of points used in the radial direction together with the maximum degree and order used in the horizontal approximation of field variables with spherical harmonics. Since five scalar fields are discretized, the total number of degrees of freedom of a simulation is $O(5N_r\ell_{\max}m_{\max})$. Namely, the triplet defining the resolution is $(144, 79, 63)$, which results in about 3.6×10^6 variables – recall that G12 has three variables, P09 and DW have one variable.

As discussed above, time in the 3D model is non-dimensional. To scale to geophysical time, one first computes the non-dimensional secular-variation timescale of the non-dipole field up to spherical harmonic degree 13, based on the average power spectra of the magnetic field and its secular variation (see Lhuillier et al. 2011b). The rescaling of the time axis is then performed under the assumption that the dynamo simulations and the Earth share the same secular-variation time scale, equal to 415 yr. With this scaling, the simulation time of the 3D model is 147 Myr; the time step is 43.09 years. The number of reversals that occur during this time frame is 109.

In contrast to the other models described above, its 3D nature makes this dynamo model amenable to quantitative comparisons against more observed properties of the Earth's magnetic field than just the axial dipole. From a morphological standpoint, the 3D model produces a magnetic field whose large-scale properties at the core-mantle boundary are in "good" agreement with well-established observations, according to the four criteria introduced by Christensen et al. (2010):

(i) the axial dipole to non-axial dipole energetic ratio;

(ii) the equatorially symmetric to antisymmetric non-dipole energetic ratio;

(iii) the zonal to non-zonal energetic ratio;

(iv) the flux concentration factor.

The terrestrial reference values for these four quantities are respectively (1.4, 1.0, 0.15, 1.5) (Christensen et al. 2010). For the 3D model, we compute average values of these quantities of,

respectively (0.72, 1.36, 0.18, 2.45). This leads to an average misfit χ^2 of 1.94, while the median value of χ^2 over the course of the numerical integration is 2.89. We refer to Christensen et al. (2010) for further details.

From a paleomagnetic perspective, it is worth noting that Sprain et al. (2019) recently introduced a method to assess the degree of spatial and temporal agreement of a simulated dynamo field with the long-term ($\sim$ 10 Myr) paleomagnetic field. The agreement is defined on the basis of five properties of the paleomagnetic field, namely the inclination anomaly, the virtual geomagnetic pole dispersion at the equator, the latitudinal variation in virtual geomagnetic pole dispersion, the normalized width of virtual dipole moment (VDM) distribution, and the dipole field reversals (in terms of the relative time spent by the dipole at transitional latitudes lower than 45°). This quantity, termed ΔQ_{PM}, is the sum of five misfits, one for each criterion. For the 3D model, we find the following values of the misfit for each criterion

$$\begin{aligned}
\Delta Q_{PM}(\text{inclination anomaly}) &= 0.84,\\
\Delta Q_{PM}(\text{equatorial dispersion}) &= 0.39,\\
\Delta Q_{PM}(\text{latitudinal dispersion}) &= 0.95,\\
\Delta Q_{PM}(\text{VDM distribution}) &= 0.81,\\
\Delta Q_{PM}(\text{reversals}) &= 1.45,
\end{aligned}$$

for a total $\Delta Q_{PM} = 4.43$. This value is good, according to Sprain et al. (2019), who argue that individual misfits lower than unity indicate an adequate similarity with the paleomagnetic field. For the study of interest here, we note that the width of the VDM distribution is adequately captured by the 3D model. On the other hand, the simulated dipole spends a fraction of time at transitional latitudes (1.2% of the model integration time) smaller than what is inferred for Earth over the last 10 Myr, which is expected to lie somewhere between 3.75% and 15.0% (see Sprain et al. 2019, for details). In summary, based on a series of metrics that have come to the fore, this

3D model compares favorably against the recent and more ancient geomagnetic field.

3.2.3 Similarities and differences across the model hierarchy

Figure 3.1 shows the axial dipole as a function of time for each model, scaled so that the average absolute value of the time series is one, and with the sign indicating polarity: a negative sign indicates today's polarity, a positive sign corresponds to a reversed polarity. The figure shows the evolution of the models' dipoles on their natural time scales described above. Each model exhibits reversals and excursions as observed in paleomagnetic reconstructions, e.g., PADM2M and Sint-2000 (Ziegler et al. 2011; Valet et al. 2005), but these events occur over different time scales and, for some of the models, the events also occur over time scales that are different from what is observed in Earth's axial dipole field. The 3D model, for example, has a reversal rate of about 0.7 reversals per Myr, but the reversal rates of the DW and P09 models are about 5 reversals per Myr, which is comparable to the average reversal rate of Earth over the last 25 Myr (Ogg 2012). In addition, the way a reversal occurs in each model can be different. Reversals of the G12 model are characterized by a continuous decay in intensity, immediately followed by a rapid increase in intensity. The DW lingers at low intensity longer than the other models (see also Figure 3.2). We also observe that the DW and 3D models, and to a lesser extent the P09 model, exhibit multiple, rapid fluctuations in sign during a reversal, see Figure 3.4. This behavior is not observed in the G12 model.

Differences between the various models can be illustrated further by comparing histograms of their intensities, shown in Figure 3.2, which also provides the models' Pearson's moment coefficient of skewness (third standardized moment, Kenney & Keeping (1966)) for the four models. It is clear that all models, except G12, are characterized by a negative skew. Thus, the P09, DW and 3D models spend more time at a lower than average intensity than at a higher than average intensity. The G12 model tends to spend more time at a higher than average intensity than at a lower than average intensity. The DW model has the smallest skew (in amplitude) and a thicker

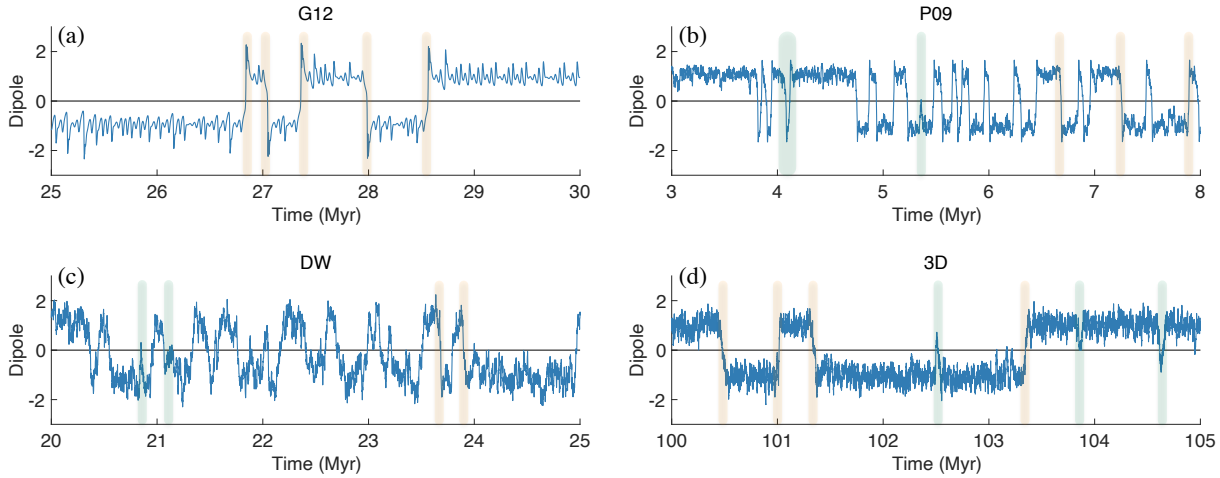

Figure 3.1: Signed dipole as a function of time for the four models considered in this study. (a) G12, (b) P09, (c) DW and (d) 3D. In each case, the amplitude is scaled so that the average absolute value of the time series is one. Some reversals and excursions are highlighted in light red and light blue.

left tail than the other three models, which indicates that it spends a considerable amount of time in low-intensity states (but other formulations of double-well models, with different parameters or even different parameterizations of the potential, may behave differently). Also shown in Figure 3.2 are histograms of the intensities of two paleomagnetic reconstructions, PADM2M and Sint-2000 (Ziegler et al. 2011; Valet et al. 2005), which document the time evolution of the virtual axial dipole moment (VADM) over the past 2 Myr at a frequency of 1 per kyr. PADM2M and Sint-2000 thus contain 2000 points and the histograms are not as well resolved as those of the four models (for which we used substantially longer simulations). This is also evident from the difference in the skewness, which is positive for Sint-2000, but negative for PADM2M. These values should be used with caution, because the estimates of skewness are contaminated by large sampling error (based on only 2000 intensities) and by the fact that low intensities (below 10%) are not present in these reconstructions. In view of the large uncertainty in the reconstructions, all four models are reasonable, at least qualitatively, i.e., in view of Figures 3.1 and 3.2, although all four models are constructed from drastically different assumptions and with very different modeling goals in mind.

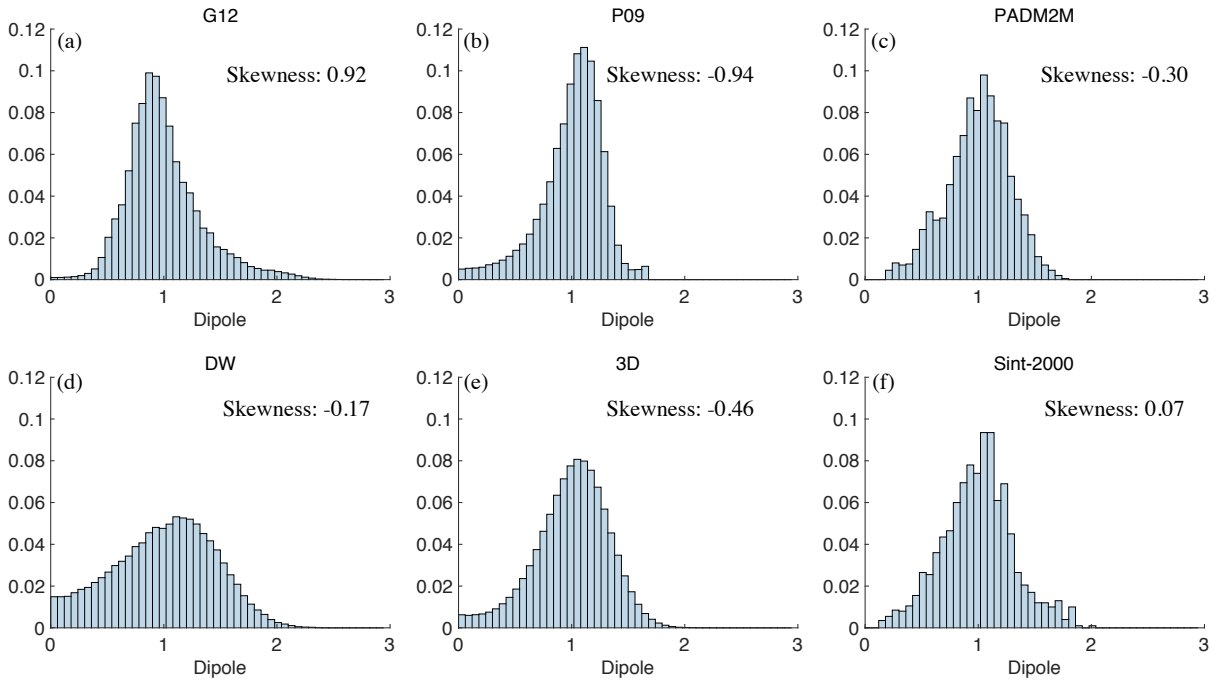

Figure 3.2: Scaled histograms of the dipole intensities of the four models and two paleomagnetic reconstructions (PADM2M and Sint-2000). (a) G12, (b) P09, (c) PADM2M, (d) DW, (e) 3D, and (f) Sint-2000. The y-axes of all histograms are scaled so that the area under the graph is equal to one and the x-axes are scaled so that one corresponds to the average intensity. Also shown is the skewness which indicates the degree of asymmetry in the distribution. A thicker tail near zero suggests that a model, or reconstruction, lingers in a state of low intensity.

3.2.4 Predictions, skill scores, and ROC curves

We want to predict whether a low-dipole event will occur within an a priori specified time interval, called the prediction horizon. We thus consider only two outcomes of an experiment. Outcome 1: yes, the event occurred during the prediction horizon; outcome 2: no, the event did not occur during the prediction horizon. As is common, we denote the outcomes of an experiment by "positives" and "negatives:"

Positive (P): the event occurred.

Negative (N): the event did not occur.

With two possible outcomes of an experiment, a prediction can result in one of four possibilities:

True positive (TP): predict that an event will occur and the event occurs.

False positive (FP): predict that an event will occur, but the event does not occur.

True negative (TN): predict that an event will not occur and the event does not occur.

False negative (FN): predict that an event will not occur, but the event occurs.

The concepts and ideas described here have been used in many areas. We make an effort to be consistent in the terminology and to bring up only the definitions we need, sticking to commonly used names (Fawcett 2006; Chicco & Jurman 2020; Joliffe 2016). For a more thorough review of the such predictions in the context of (medical) imaging, see Barrett & Myers (2003), Chapter 13, where, a prediction strategy of the type discussed here is called a binary decision. In the language of machine learning, the problem of predicting "an event will occur within the horizon" or "no event will occur within the horizon" is called a classification problem, similar to distinguishing dogs from cats (Goodfellow et al. 2016).

It is clear that a good prediction strategy should be characterized by a large number of

true positives and true negatives, but a small number of false positives or false negatives. A skill score is a quantitative means for describing the quality of a prediction strategy. There is a large number of skill scores and these usually require that one applies the prediction strategy, say n times, followed by counting the number P of positives that occurred, the number N of negatives that occurred, and the true/false positives and true/false negatives. Which of the many skill scores is most appropriate depends on the problem one wishes to solve. For example, one may define the accuracy by

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}. \tag{3.5}$$

A good prediction strategy should be characterized by a high accuracy, but a bad prediction strategy may also be characterized by a high accuracy. For example, the event "a low-dipole event occurs within the prediction horizon" is rare compared to the event "no low-dipole event occurs within the prediction horizon" (unless the prediction horizon is large). This means that the prediction strategy "predict that no low-dipole event occurs within the prediction horizon" is characterized by a high accuracy, but this strategy is useless because it can never achieve a true positive (the event "a low-dipole event occurs within the horizon" is never predicted). Other skill scores, e.g., the F_1 score

$$\mathrm{F}_1 = \frac{2\,\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{3.6}$$

the critical success index (CSI)

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{3.7}$$

or Matthew's correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{3.8}$$

are designed to alleviate these issues and are applicable in problems where the occurrence of the event is rare.

One can also compute the true-positive-rate (TPR) and false-positive-rate (FPR), defined by:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}}. \tag{3.9}$$

TPR and FPR have the desirable property that they are independent of the frequency of the event and a good prediction strategy should have a high TPR and low FPR, independently of how often an event occurs. Ideally, $\mathrm{TPR} = 1$, so that all events are predicted correctly, and $\mathrm{FPR} = 0$, so that no false positives occur.

The bulk of this paper is concerned with predictions based on whether the dipole is below a specified threshold. Naturally, one may investigate how the skill of the predictions depends on the threshold. For example, one can compute skill scores as functions of a varying threshold and determine an optimal threshold as the one that maximizes the skill score. One can also compute the TPR and FPR as functions of the threshold. The line that a varying threshold traces out in $\mathrm{TPR} - \mathrm{FPR}$ space is called the receiver operating characteristic (ROC) curve. Three examples of ROC curves are illustrated in Figure 3.3. The figure also shows the chance line, defined by a straight line at a 45° angle, on which the (FTR,TPR) points should lie when randomly guessing occurence of events, and varying the probability with which one makes this guess.

The ROC curve of a good prediction strategy should be above the chance line and should quickly transition from the origin towards $(0, 1)$, thus being characterized by a high TPR and a small FPR. One can use ROC curves as qualitative tools to assess different prediction strategies. We have included labels for the three ROC curves in Figure 3.3, that identify which strategies are good, worse or bad.

We note that, besides an impressively large body of work across many disciplines, it remains difficult to unambiguously argue that a prediction strategy is good or bad, or if one prediction strategy is better than another. As a simple example, consider the green ROC curve in Figure 3.3 with threshold levels labeled by A and B. It is not easy to say which threshold level one should choose. Threshold A leads to the smallest FPR while also achieving $\mathrm{TPR} = 1$, but

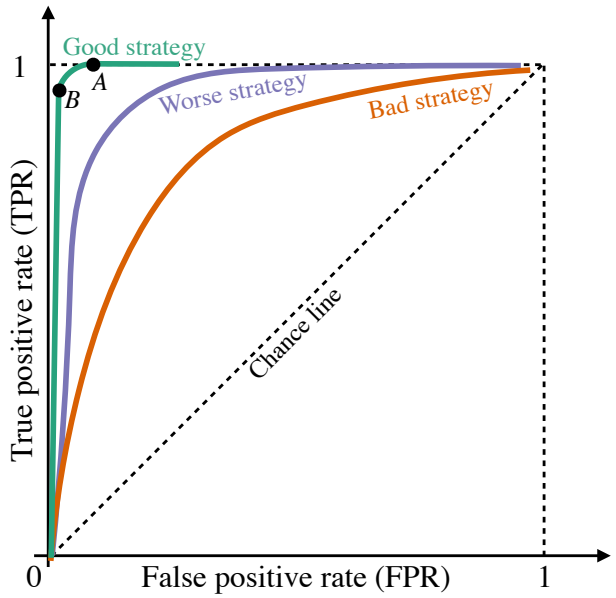

Figure 3.3: Three examples of ROC curves. Two threshold levels, labeled *A* and *B*, are identified on one of the curves (see text for the definition of the chance line).

threshold *B* achieves a smaller FPR than threshold *A*, at the cost of a slightly smaller TPR. The difficulties arise because many issues, such as how dangerous false positives are compared to false negatives, are problem dependent and remain subjective.

3.3 Finding thresholds for the prediction of low-dipole events

In simple terms, our prediction strategy is

If the dipole intensity drops below a threshold, then the field will continue to decay and a low-dipole event will occur within the prediction horizon.

The situation, however, is more delicate because the models exhibit complex behavior while undergoing reversals or major excursions. The subtleties can be illustrated by considering the excerpt of the 3D simulation shown in Figure 3.4, where we highlight reversals and major excursions during which the axial dipole field temporarily changed sign. One wishes to define a reversal event as the transition of a *strong* dipole field in one polarity, to a *strong* dipole field in

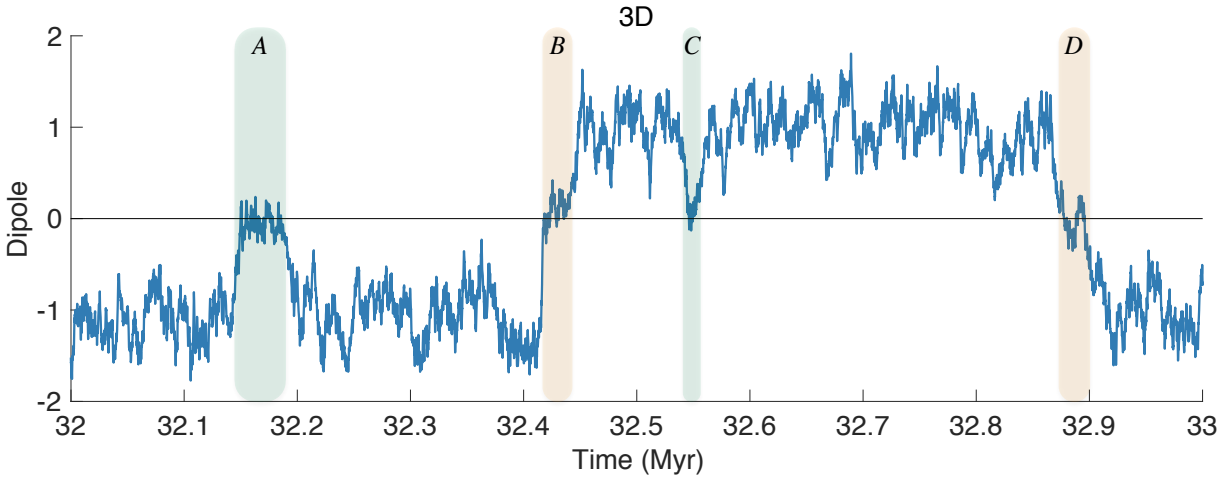

Figure 3.4: Excerpt of the 3D simulation showing the signed dipole as a function of time (same scaling as in Figure 3.1). Four events are labeled $A - D$.

the opposite polarity, rather than just a short-term temporary change of polarity. In fact, the field may quickly change polarity several times while the field is weak when undergoing a reversal. This occurs in the 3D simulation during the events labeled B and D in Figure 3.4. Similarly, one wishes to interpret event A (or C) as a single major excursion, rather than a sequence of reversals. We thus revise the simple prediction strategy above to ensure that each reversal or major excursion, labeled by $A - D$ in Figure 3.4, is considered as a single low-dipole event. A careful definition of low-dipole events is provided below.

3.3.1 Precise formulation of threshold-based predictions

We introduce definitions that allow us to clearly specify the events and predictions whose skill we want to study. We start with the definition of the event we want to predict.

Definition: *Low-dipole event*. A low-dipole event starts when the intensity drops below a specified value, called the *start-of-event threshold* (ST), or if a the dipole changes its sign[2], and ends when the intensity exceeds a second specified value, called the *end-of-event threshold* (ET).

[2]The addition of the "or-statement" is relevant only in the context of paleomagnetic reconstructions, see Section 3.5.

The *event duration* is the time interval from start to end of the event.

This definition ensures that a low-dipole event describes reversals and major excursions, because the field may drop below the ST, but can build back up above the ET without changing polarity. Events *A* and *C* in Figure 3.4 are examples of this situation. We also emphasize that the event duration is defined implicitly by the start and end of an event and may vary considerably across several low-dipole events. In Figure 3.4, for example, the low-dipole event *A* has a much larger event duration than event *C*, but both events are major excursions.

To define a strategy for predicting low-dipole events, we introduce the *prediction horizon* (PH), which is the time window during which we predict that a low-dipole event will start to occur. Note that we do not make any prediction as to when precisely the low-dipole event starts – we merely predict that a low-dipole event will start (or not) at some point during the PH. We further make no predictions as to when the low-dipole event will end. With the above definitions, the prediction strategy can be stated precisely.

Definition: *Threshold-based predictions for low-dipole events*. We predict that a low-dipole event will start to occur within the prediction horizon if the intensity drops below a *warning threshold* (WT); we predict that no low-dipole event will start during the prediction horizon if the intensity is above the WT; we stop making predictions from the time the low-dipole event started (intensity below ST) until the event ends (intensity above ET).

We make no predictions while the event is observed, because a prediction made while an event is happening is of limited use. Our prediction strategy is illustrated in Figure 3.5. In this illustration, ST and ET are chosen such that Event *A* is a single event; fast oscillations in polarity occur while the field is weak. The prediction strategy leads to TNs followed by FNs and TPs for Events A and B. The false negatives occur because, given the prediction horizon and average intensity, the warning threshold is small, so that the events tend to be predicted a little too late. The figure also illustrates FPs, which occur when the field drops below the WT, but no low-dipole event occurs

because the field does not continue to drop below the ST. True negatives occur often, because reversals and low dipole events are rare. In the figure, TNs occur whenever the "Truth" (bottom) and the prediction (center) are both at zero.

Finally, note that with our definitions, one may require that

$$\mathrm{ST} < \mathrm{WT} < \mathrm{ET}, \tag{3.10}$$

because other choices for WT lead to rather strange prediction strategies. If ET $\leq$ WT, for example, then a low-dipole event would be predicted immediately after an event just ended.

3.3.2 Scaling thresholds and re-scaling time

For each model, we define all three thresholds (warning, start-of-event, and end-of-event thresholds) as a fraction of the average intensity of the model. Such a scaling of the thresholds makes comparisons across the hierarchy of models easier to understand. For the rest of this paper we fix the start-of-event and end-of-event thresholds as: ST =10% and ET=80%. With these choices, we focus on events that start when the intensity is very low and which end when the field has nearly fully recovered (see Figure 3.4). The choice of ST =10% is guided by the consideration that we want to focus on events that correspond to reversals and major excursion. During a reversal, the signed dipole can reach an arbitrarily low value, before switching sign. During a major excursion, the dipole amplitude is very low, but we do not necessarily observe a switch in the sign. Moreover, paleomagnetic reconstructions, such as PADM2M and Sint-2000 (Ziegler et al. 2011; Valet et al. 2005), have difficulties with resolving small dipole values. The paleomagnetic reconstructions we consider below consist of signed Virtual Axial Dipole Moments (VADM), which are proxies for the true axial dipole magnitude. The weakest VADMs recorded are about 10-20% of the present axial dipole field intensity (see, e.g., Constable & Korte 2006; Hulot et al. 2010a). This is caused by (i) VADM reconstructions sensing the non-dipole field during

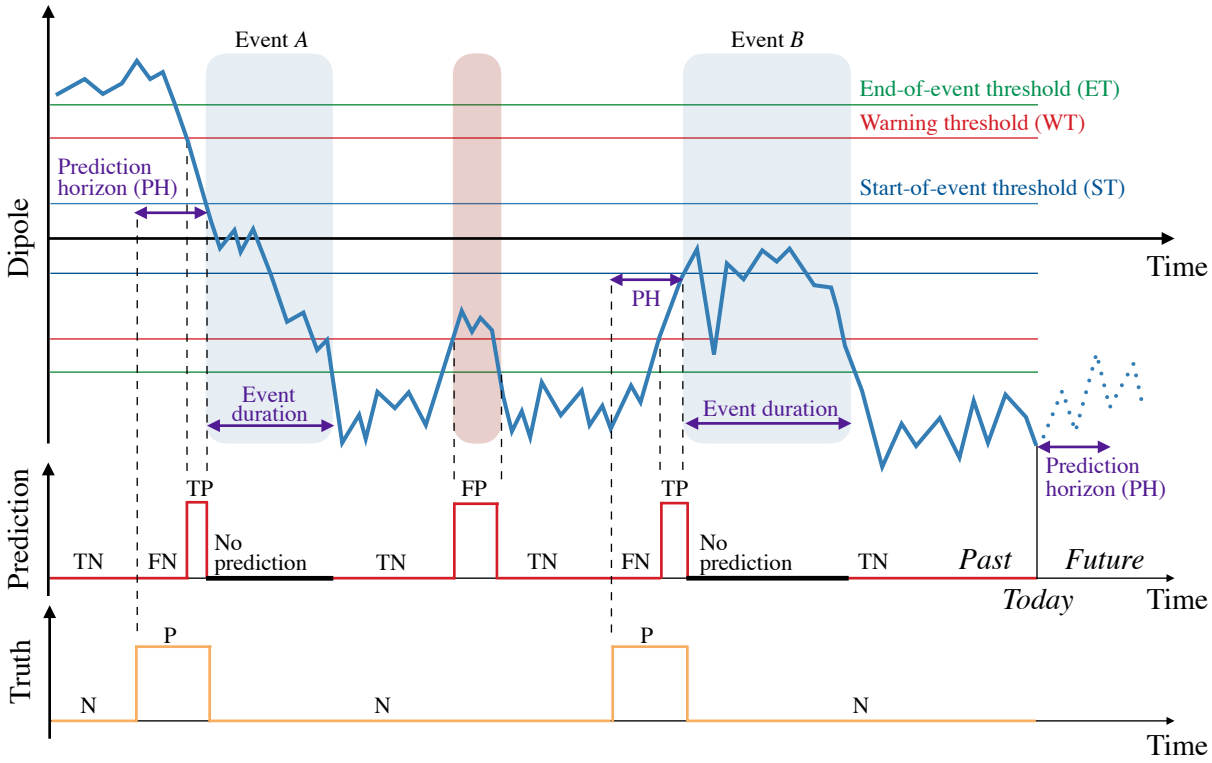

Figure 3.5: Illustration of the prediction strategy. *Top*: dipole (solid blue) as a function of time. The thin blue, green and red horizontal lines represent the start-of-event, the end-of-event and the warning thresholds. Two low-dipole events are labeled *A* (reversal) and *B* (excursion), and we indicate their event durations. Highlighted in red is a period of low intensity, which is not a low-dipole event, but where the low intensity causes false positives (FP). Towards the right, we illustrate a prediction over a given prediction horizon, which will lead to true negatives (TN). The prediction horizon also defines the true labels, (see bottom panel). *Center*: prediction as a function of time. The red line at zero corresponds to the prediction "no low-dipole event occurs during the prediction horizon," and the red line at one corresponds to the prediction "a low-dipole event occurs during the prediction horizon." The thick black line segments correspond to periods during which no prediction is made. For events *A* and *B*, we first observe TNs, followed by false negatives (FN), caused by the warning threshold being small; then we observe TPs followed by a period during which no prediction is made. *Bottom*: true occurrences of low-dipole events within the prediction horizon. The orange line at zero corresponds to negatives (N), i.e., "no low-dipole event occurs during the prediction horizon." The orange line at one corresponds to positives (P), i.e., "a low-dipole event occurs during the prediction horizon."

a low-dipole event; (ii) VADM reconstructions are temporally filtered by sediment recording processes; and (iii) additional smoothing is introduced by modeling choices and stacking of the relative paleointensity (RPI) records (some of the individual records may have a higher resolution and features that are not aligned in time are smoothed out). As we will see, by choosing ST =10%, we ensure that only events that experienced at least one temporary change of sign in the axial dipole are considered as events of interest within PADM2M and Sint-2000 (see Section 3.5).

Nonetheless, the precise values of ST and ET are not critical because our overall approach is robust with respect to choices. This is evident from a limited number of numerical experiments we performed with different choices of ET and ST. Specifically, we tried the combinations ST=10% and ET = 50%, ST=20% and ET = 50%, and ST =20% and ET=80% and obtained qualitatively and quantitatively similar results.

We rescale time in each model so that the prediction results are comparable across the hierarchy of models. A natural choice for this time scale is the *average event duration* (AED). That is, we compute the average event duration given the natural time-scale of each model, and then rescale time so that one time unit corresponds to one average event duration. The average event duration (AED) for each of the models is listed in Table 3.1. For the simplified models (G12, P09 and DW), the statistics of the event duration are computed from simulations that include about 550 events. For the 3D model, we use the entire duration of the simulation to compute the statistics of the event duration.

The prediction horizon is defined as a fraction of the average event duration. We focus on the prediction horizon PH $= 1\times$ AED, i.e., we focus on short-term predictions of low-dipole events, attempting to predict whether a low-dipole event will start with a lead time comparable to the event's duration. We also consider prediction horizons of $0.5\times$ AED or $1.5\times$ AED, to show how the prediction skill degrades with longer prediction horizons, but also to demonstrate the robustness of our approach (which is not sensitive to minor variations of the various parameters).

Table 3.1: This table summarizes key results obtained throughout the paper. We list all information in this one table to make it easier to make connections between the various quantities listed. Description of each column. *First column*: the model or paleomagnetic reconstruction considered. *Second column*: number of low-dipole events in the verification portion of a simulation/ paleomagnetic reconstruction. Values in brackets are the number of events in the training data. *Third column*: maximum MCC (prediction skill) achieved for optimal WT (see Section 3.2.4 for the definition of MCC). Values in brackets are for training data. Verification and training data are explained in Section 3.4.2 for the models and in Section 3.5.2 for the paleomagnetic reconstructions. *Fourth column*: optimal WT that maximizes MCC over the training data (see Section 3.3.3). *Fifth column*: average duration of a low-dipole event (AED). Values in brackets are standard deviations. *Sixth column*: average decay time (ADT) with standard deviations in brackets. See sections 3.4.1 (models) and 3.5.1 (paleomagnetic reconstructions) for definitions of average event duration and decay time and their computation. *Seventh column*: ratio of average decay time to average event duration. All results listed here correspond to a prediction horizon PH $= 1$, a start-of-event threshold ST $= 10\%$, and an end-of-event threshold ET $= 80\%$.

	# of events	MCC	$\hat{\text{WT}}$	Event duration (AED)	Decay time (ADT)	$\rho = \frac{\text{ADT}}{\text{AED}}$
G12	554 (5)	0.96 (0.97)	30.75%	3.2 kyr (0.1 kyr)	26.9 kyr (2.0 kyr)	8.49
P09	551 (5)	0.57 (0.56)	54.50%	6.0 kyr (3.9 kyr)	10.8 kyr (5.0 kyr)	1.81
DW	551 (5)	0.31 (0.30)	69.25%	16.1 kyr (12.2 kyr)	8.5 kyr (4.7 kyr)	0.53
3D	368 (5)	0.12 (0.14)	17.50%	16.4 kyr (11.6 kyr)	5.7 kyr (4.2 kyr)	0.35
PADM2M	2 (4)	0.62 (0.73)	50.75%	11.7 kyr (8.1 kyr)	25.5 kyr (10.9 kyr)	2.19
Sint-2000	2 (4)	0.44 (0.77)	36.75 %	10.2 kyr (8.7 kyr)	32.0 kyr (15.1 kyr)	3.15

3.3.3 Finding thresholds via maximization of skill scores

For a fixed prediction horizon, we compute an optimal warning threshold as follows. For a given dipole time series, we compute a skill score for varying WTs, using a regular grid with spacing of 0.25%. The WT that leads to the largest skill score is selected as the optimal WT: $\hat{\mathrm{WT}} = \arg\max \mathrm{Skill}(\mathrm{WT})$.

This approach can be implemented with a variety of skill scores, e.g., MCC, CSI or F_1. For the short prediction horizons we consider, we did not notice any significant differences in the optimal WTs one finds regardless of which skill score is used, with the exception of the ACC score, which is not robust with respect to imbalances in the data (one event occurring more often than the other). Below we present results obtained by using MCC, because it recently has been reported to be more appropriate than the F_1 score for binary classification (Chicco & Jurman 2020), but other skill scores (not ACC) may be used to obtain similar results.

To prevent overfitting, it is necessary to validate a prediction strategy by applying it to an independent data set. An optimal WT is determined by using a given dipole time series, which we call the training data set. The optimal WT is then applied to an independent time series, which we call the verification data set, and the skill score is computed for the verification data. For the simplified models (G12, P09, DW), the verification data are independent simulations (using different initial conditions in the case of the deterministic G12 model and different initial conditions and random forcing in the case of the stochastic P09 and DW models). For the 3D model, we compute the optimal WT by using only a portion of the simulation as training data, and then use the remainder of the simulation as verification data.

3.4 Application to a hierarchy of models

We apply threshold-based predictions to the models in the hierarchy. For each model, we predict a low-dipole event about as far ahead of time as one expects the event will last. In

our terminology, this means that the prediction horizon is equal to one average event duration (PH $= 1\times$ AED), but we also consider slightly longer ($1.5\times$) and slightly shorter ($0.5\times$) PHs. We also test if useful threshold-based predictions can be made if the training period is short and, therefore, contains only a small number of low-dipole events.

3.4.1 Skill of threshold-based predictions

Qualitative comparison and illustration

We first qualitatively assess threshold-based predictions by inspection of ROC curves. The ROC curves shown in Figure 3.6 are computed using the entire model run in the case of the 3D model, and long simulations with around 550 events for the G12, P09 and DW models, see Section 3.3.2. We note that, for all four models, the ROC curves get closer to the chance line (higher false positive rate, lower true positive rate) as the prediction horizon increases. This implies that the predictions get worse, by any measure, as the prediction horizon increases. This means, perhaps not so surprisingly, that predictions via a threshold-based strategy are more difficult to do when the prediction horizon is large. More interestingly, we note that for any fixed PH, the ROC curves of the G12 or P09 models are further from the chance line than the ROC curves of the DW and 3D models. This suggests a "ranking" of the models in terms of how skillful threshold-based predictions are [3]. We investigate this ranking quantitatively via MCC skill scores below.

For each model, we illustrate threshold-based predictions for which an optimal warning threshold is found by maximizing MCC skill score, as a function of the warning threshold. The data sets used for finding the optimal WTs are the entire model run in the case of the 3D model, and long simulations with around 550 events for the G12, P09 and DW models, see Section 3.3.2 (no distinction between training and verification data). This results in optimal WTs

[3] But a higher ranking in predictive skill does not imply that the model is "better," i.e., more similar to Earth's axial dipole, see Section 3.5.

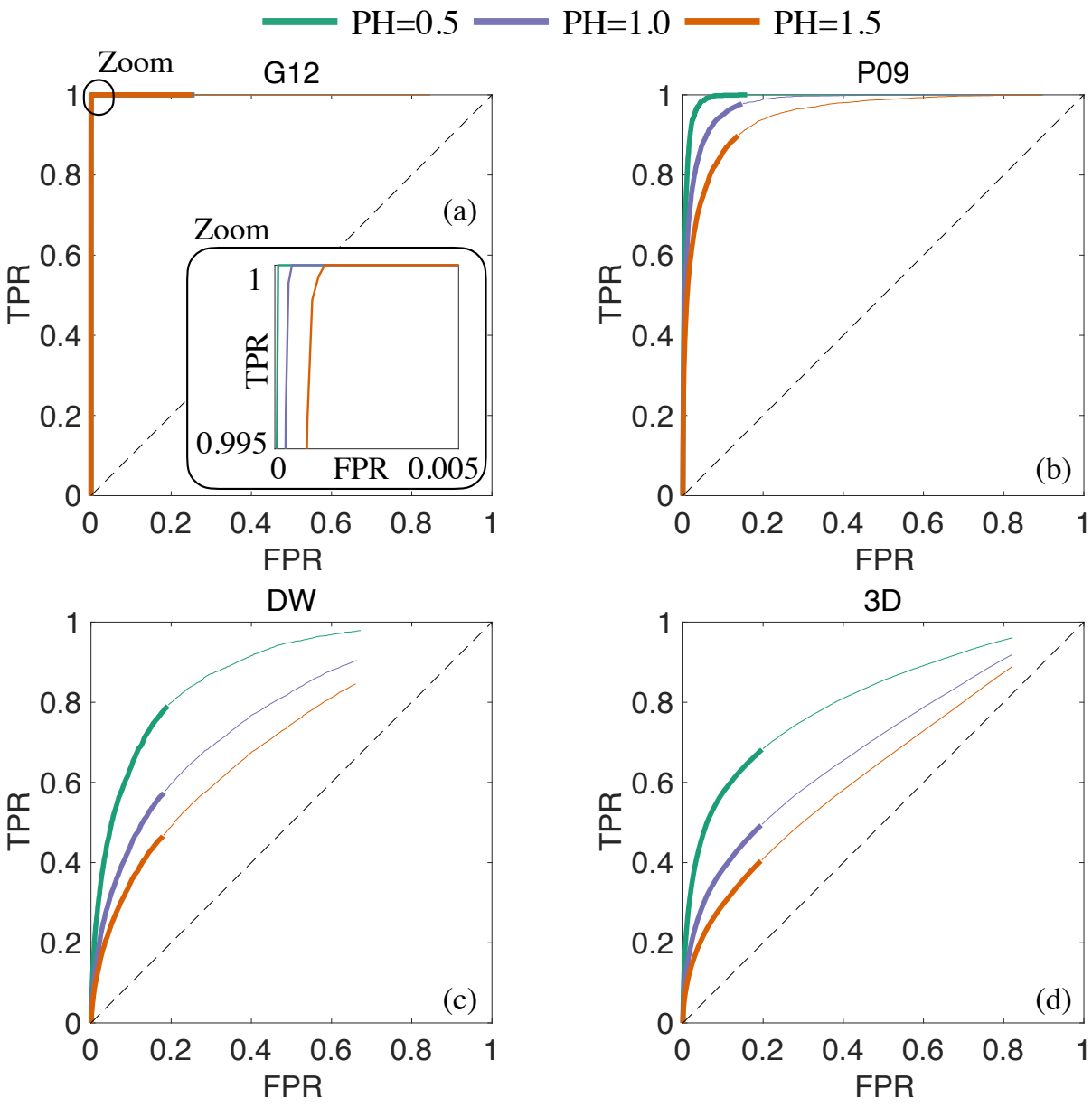

Figure 3.6: ROC curves for the four models and three prediction horizons, with PH = 0.5 in green, PH = 1 in purple, and PH = 1.5 in orange. (a) G12, (b) P09, (c) DW, (d) 3D. A ROC curve is the collection of TPR/FPR pairs one obtains when varying the warning threshold. The thicker line corresponds to TPR/FPR pairs for which ST $<$ WT $<$ ET. The thin lines continue the ROC curves for WT $\geq$ ET. The figure-in-figure in (a) (ROC curves for G12) shows a zoom near the (0,1) point to illustrate that the three ROC curves, corresponding to different PHs, do not overlap. The ROC curves are computed using long simulations, containing a large number of low-dipole events (see text for details).

of $\hat{\mathrm{WT}}_{\mathrm{G12}} = 31.25\%$, $\hat{\mathrm{WT}}_{\mathrm{P09}} = 43.00\%$, $\hat{\mathrm{WT}}_{\mathrm{DW}} = 60.25\%$ and $\hat{\mathrm{WT}}_{\mathrm{3D}} = 45.50\%$ for respectively the G12, P09, DW and 3D models. Results for a prediction horizon PH $= 1$ are shown in Figure 3.7; results for PH $= 0.5$ or PH $= 1.5$ are similar. We plot excerpts of the dipole time series of the four models, along with two graphs that illustrate the predictions and their validity. Each model is represented by one sub-figure which contains three panels. The top panel shows an excerpt of the dipole time series. We show a time interval of 50 non-dimensional time units for each model and each model exhibits two events during this time interval (recall that time is scaled by the average event duration, AED, see Table 3.1). The orange lines in the bottom of each sub-figure are zero if no low-dipole event starts within the prediction horizon, and they are one if a low-dipole event starts during the prediction horizon. Because we show the same time-interval in non-dimensional units, the intervals during which the orange line is at one are of equal width across all four sub-figures. The red lines in the center panels are zero if no low-dipole event is predicted to start during the prediction horizon; the lines are one if a low-dipole event is predicted to start during the prediction horizon. Thus, the overlap of the red and orange lines defines TPs, FPs, TNs and FNs, and a large overlap corresponds to a skillful prediction. For example, a FP corresponds to a situation where the orange is at zero while the red line is at one; a FN corresponds to a situation where the orange line is at one while the red line is at zero.

We note that predictions for the G12 model lead to a small number of false positives or false negatives. In the excerpt shown for the G12 model in Figure 3.7, there is only one false positive, caused by the prediction starting one time step too early (during the first of the two events shown). For the DW and 3D models, on the other hand, we note a large number of false positives and false negatives, which renders threshold-based predictions unreliable for these models. Comparisons of the graphs for G12 and the DW and 3D models suggest that threshold-based predictions for the DW or 3D model are indeed worse, by any measure, than those for the G12 model. In the case of P09, we note a larger number of false positives and false negatives than in the case of G12, but false positives or false negatives occur less frequently

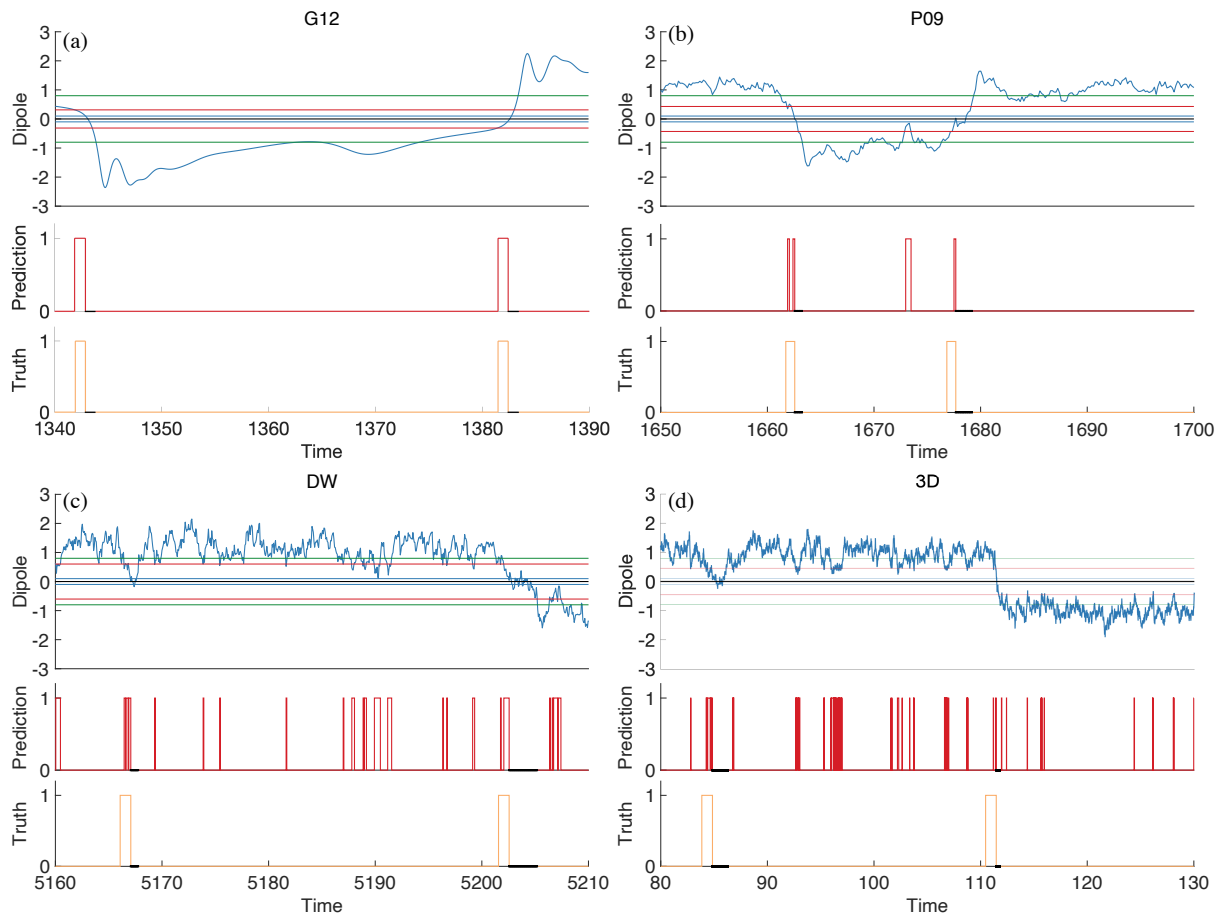

Figure 3.7: Illustration of threshold-based predictions for the four models. (a) G12, (b) P09, (c) DW, (d) 3D. The plots show predictions over a time period of 50 dimensionless time units and for a prediction horizon of PH = 1. The corresponding (optimal) warning thresholds (expressed in percent of the average intensity) are $\hat{\text{WT}}_{\text{G12}} = 31.25\%$, $\hat{\text{WT}}_{\text{P09}} = 43.00\%$, $\hat{\text{WT}}_{\text{DW}} = 60.25\%$ and $\hat{\text{WT}}_{\text{3D}} = 45.50\%$ for respectively the G12, P09, DW and 3D models. These optimal WTs are computed using time series of the four models that contain a large number of events (see text for details). For each model, a dimensionless time is defined via a scaling of time with the average event duration (see text for details). Each sub-figure contains three panels. *Top*. Blue: dipole time series. Blue/green/red horizontal lines: start-of-event/end-of-event/warning thresholds. *Center*. Graphs are zero if the threshold-based prediction is "no low-dipole event will start during the prediction horizon;" graphs are one if the threshold-based prediction is "a low-dipole event will start during the prediction horizon." *Bottom*. The graphs are zero if no low-dipole event starts within the prediction horizon; the graphs are one if a low-dipole event starts during the prediction horizon. Black lines in the center and bottom graphs denote times when no predictions are being made.

than for the DW or 3D models. Consistent with what was suggested by Figure 3.6, the skill of threshold-based predictions for the P09 model thus seems to fall in between the skills of threshold-based predictions for the G12 (very high skill) and DW/3D models (very low skills).

Quantitative comparison and ranking

We compute MCC skill scores to quantitatively compare the skill of threshold-based predictions for the various models. To avoid over-fitting we now compute skill scores on verification data, i.e., data that are *not* used for computing the optimal warning threshold, as described in Section 3.3.3.

We generate training and verification data as follows. For the G12, P09 and DW models, the training data are the long simulations that were also used in Section 3.3.2. The verification data are ten independent simulations, each of length 10^4. For the 3D model, we create training and verification data by "chopping up" the overall simulation as follows. We split the simulations into two parts of equal length and use one for training and the other for verification. We then repeat the procedure, but split the simulation into three equally long portions, using one for training and two for verification. Finally, we split the simulation into four equally long portions, and use one for training and three for verification. This procedure leads to six MCC scores over verification data. Generating multiple verification data sets in this way allows us to estimate the variability in the skill of threshold-based predictions for all four models.

Results are shown in Figure 3.8, where we plot MCC scores for the four models for threshold-based predictions with prediction horizon PH=1. We only show the results for a prediction horizon PH $= 1$, but one obtains qualitatively the same results with PH $= 0.5$ or PH $= 1.5$. We note that the variation in skill over the different verification data sets is small. This suggests that the verification data sets are "large enough" so that variation in the verification data does not affect the scores. Moreover, our results confirm the ranking of the skill of threshold-based predictions that we anticipated from inspection of ROC curves. Specifically, we rank the models

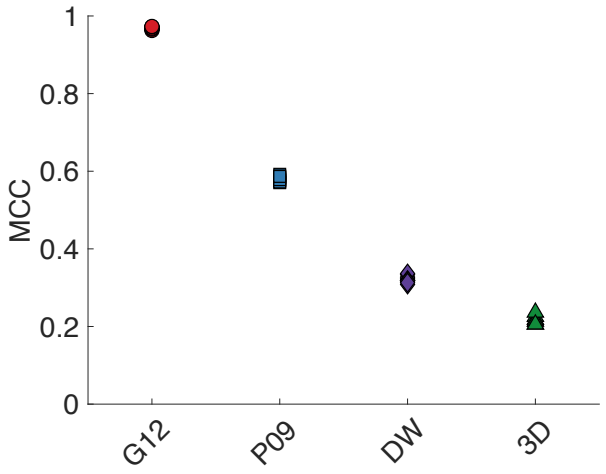

Figure 3.8: MCC skill scores (verification) of the four models for prediction horizon PH = 1. For each model, several MCC scores are shown. The MCC scores are computed over multiple sets of verification data. The training data are long simulations containing many low-dipole events (see text for details).

(skill from high to low) in terms of their predictability via intensity thresholding as: G12, P09, DW and 3D. Indeed, we found that this result is independent of the choice of skill score – one obtains qualitatively and, to a large extent, quantitatively the same results using, e.g., the F_1 or CSI skill scores.

This ranking and, more generally, the skill of threshold-based predictions appears to be determined by an interplay of:

(i) *The extent of variation in the dipole intensity*: a high potential for false positives results if the intensity dips to low values regularly, but if no low-dipole event follows.

(ii) *The decay rate prior to a low-dipole event*: a quick decay results in a high potential for false negatives.

For (i), we recall the intensity histograms of Figure 3.2, which show that the P09, DW and 3D models (low skill) spend more time at low intensity values than the G12 model (high skill). For (ii), we compute the *average decay time* (ADT), which measures how quickly the dipole intensity decays prior to a low-dipole event. We define the *decay time* as the absolute value of the

time difference between the start of the event and the last previous instance at which the field exceeded the end-of-event threshold (ET, 80% of its average value). We list the ADT for the four models in Table 3.1, with standard deviations. These ADT should be compared to the average event durations (AED), also listed in Table 3.1. Recall that the event duration is defined by the time interval that starts when the dipole intensity drops below a given start-of-event threshold (10% of the average value) and ends when the dipole intensity exceeds a given end-of-event threshold (80% of the average value). Thus, the average event duration describes how quickly on average the dipole recovers to a large value after it dropped to a low value. If the average decay time is larger than the average event duration (ADT $>$ AED), then low-dipole events occur slowly; if the average decay time is smaller than the average event duration (ADT $<$ AED), then low-dipole events occur quickly. The two behaviors are illustrated by the G12 and 3D models in Figures 3.9(b) and 3.9(c).

In Figure 3.9(a), we plot the ratio of the average decay time to the average event duration for the four models (see Table 3.1). For brevity, we introduce an abbreviation for this ratio:

$$\rho = \frac{\text{ADT}}{\text{AED}}. \tag{3.11}$$

We note that ρ follows a similar trend as the MCC score. In particular, the ranking of the models it leads to is identical to the ranking inferred from the MCC skill score. This suggests that the skill of threshold-based predictions is influenced by how quickly low-dipole events occur with respect to their duration. If they occur slowly (ADT $>$ AED, $\rho > 1$), then threshold-based predictions have a high skill. If they occur quickly (ADT $<$ AED, $\rho < 1$), then threshold-based predictions may have a low skill.

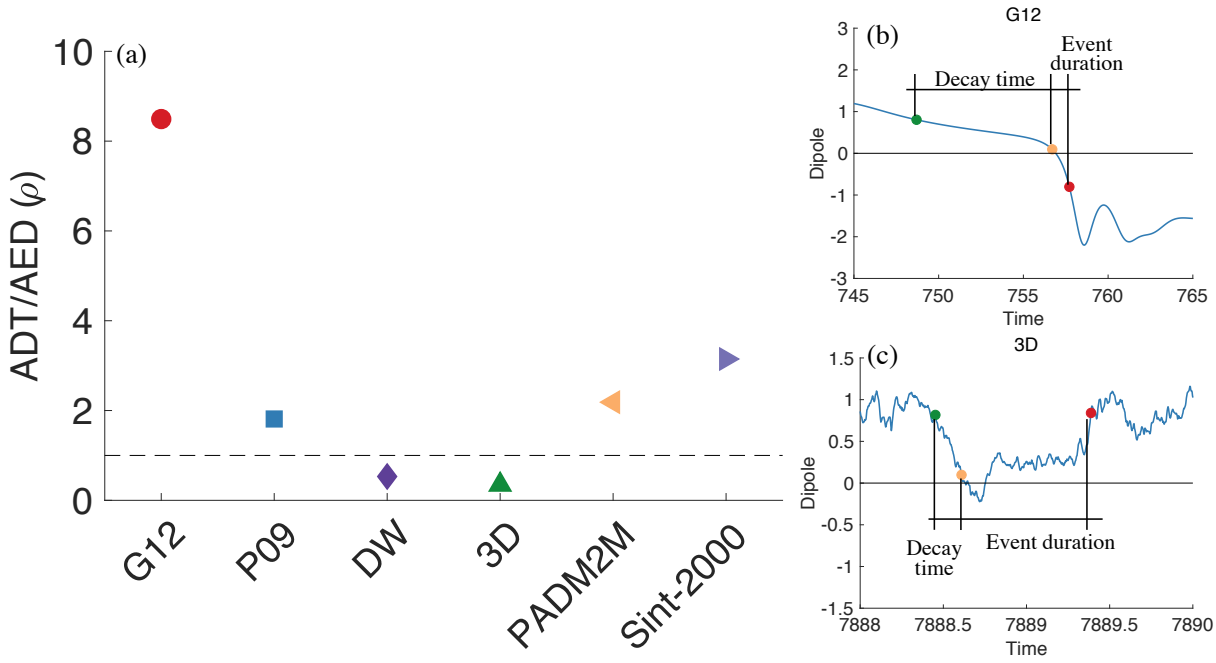

Figure 3.9: (a): ratio ρ of the average decay time (ADT) to the average event duration (AED) for the four models and two paleomagnetic reconstructions (PADM2M and Sint-2000, see Section 3.5). Also shown is the $\rho = 1$ line (dashed). (b): illustration of the decay time and event duration of an event for G12. (c): illustration of the decay time and event duration of an event for the 3D model. The beginning of the decay is marked in green, the start of an event is marked in orange and the end of an event is marked in red. The decay time is the time interval between the start of the decay and the start of an event. The event duration is the time interval between the start and end of an event.

3.4.2 Robustness of skill to a short training period

Motivated by the fact that the observational record is short (the PADM2M and Sint-2000 reconstructions that we investigate in Section 3.5 extend over 2 Myr and contain only six low-dipole events), we investigate the robustness of the optimal WT and corresponding skill with respect to the duration of the training data. For each model we compute an optimal WT for several training data sets of different durations, and, hence, containing a different number of events. Results for a prediction horizon PH $= 1$ are shown in Figure 3.10(a). For G12, we note that the optimal WT is nearly independent of the duration of the training data set. This means that, for this model, one can find a useful warning threshold from a rather short training period. For all other models, we observe a variation of the optimal warning threshold as we vary the duration of the training period. The variations are most significant for the DW model, for which the optimal WT varies from about 25% to nearly 80% (which is the maximum allowed value). For P09, the optimal WT varies between 35% and 55%, but there seems to be a plateau of nearly constant WT for training data sets that contain 20-35 events. For the 3D model, we observe a variation of the optimal WT between 20% to about 40%. Again, we note a plateau of nearly constant WT for training data with 10-40 events.

Variations in the optimal WT, however, do not necessarily imply variations in the resulting MCC skill score. This is shown in Figure 3.10(b). Here, we use the optimal WTs obtained from the same various training periods (and shown in Figure 3.10(a)), but compute the MCC over the verification data. For the simple models (G12, P09 and DW), the verification data are the long simulations (with about 550 events, see Section 3.3.2). For the 3D model, the verification data are the portion of the simulation that was not used during training.

We observe that the MCC skill score of threshold-based predictions is nearly independent of the duration of the training data. This is consistent across the hierarchy of models and suggests that the shortness of the observational record may not be the critical limiting factor for determining a useful warning threshold. Our numerical results indeed suggest that a useful WT can be found

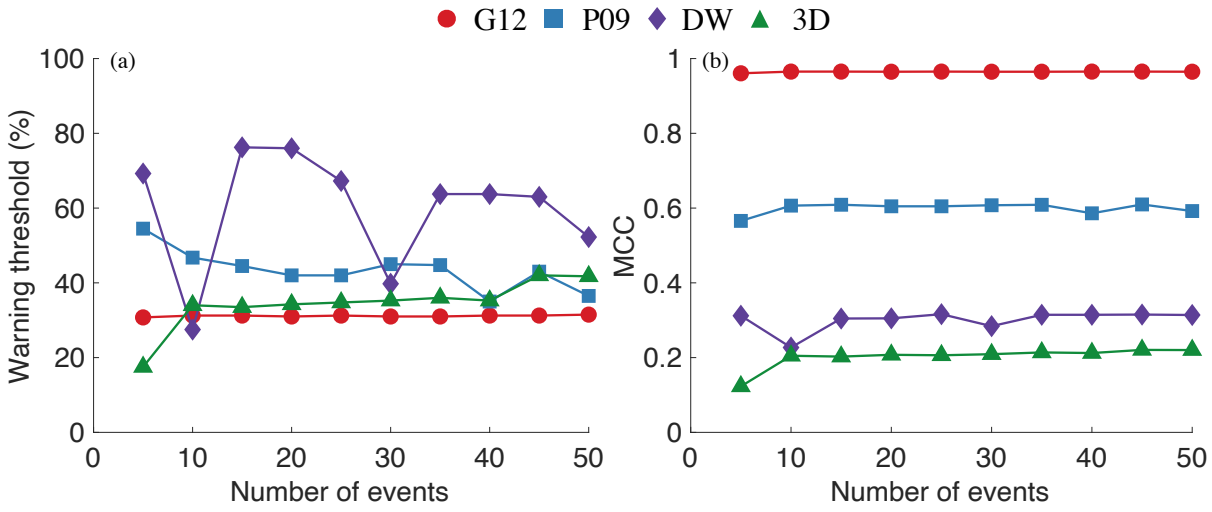

Figure 3.10: (a) optimal warning threshold as a function of the number of events contained in the training data. (b) MCC computed over verification data as a function of the number of events contained in the training data. The prediction horizon is PH = 1.

even if the training period is short and comparable with the observational record.

The reason *why* the skill is independent of the duration of the training data varies across the hierarchy. This can be understood by considering how MCC depends on WT, which we compute and show in Figure 3.11. If the MCC vs. WT graph is sharply peaked around an optimal value, and if the peak is nearly independent of the duration of the training period, then a good WT can be found even with a limited amount of training data. This is the case for the G12 model. If the graph of MCC skill score plateaus for large values of WT, then rather different WT values can produce a similar skill scores. This is the explanation for why drastic variations in optimal WT cause nearly no variations in the resulting optimal MCC in case of the DW model in Figure 3.10.

3.4.3 Impact of data filtering

Threshold-based predictions for the 3D and DW model have a low skill compared to P09 or G12. This could be due to the quick changes in polarity that we observe in the 3D and DW models, and that occur on short time scales (recall Figure 3.1). These are absent from the P09 or G12 models. Palomagnetic reconstructions such as PADM2M and Sint-2000, which are

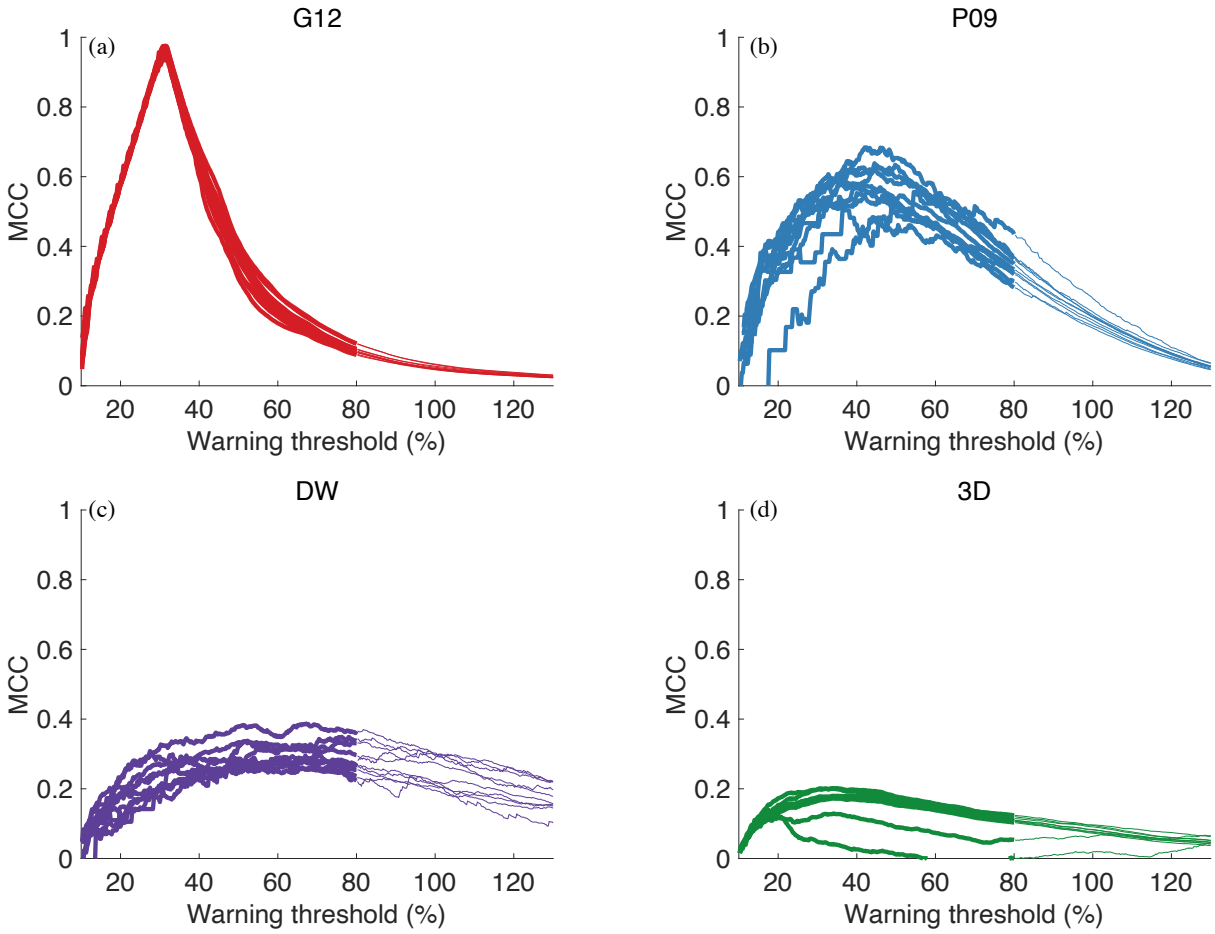

Figure 3.11: MCC skill score as a function of WT for the four models. (a) G12, (b) P09, (c) DW, (d) 3D. The various graphs shown for each model differ in the number of events contained in the training data (see text for details). The thin lines continue the curves for WT $\geq$ ET.

inherently smoothing the field they record through the slowly depositing sedimentary process, also fail to show such a behavior. One may thus wonder if the 3D or DW models could become more amenable to threshold-based predictions if the dipole is smoothed in an analogous way.

To test this possibility, we first consider the 3D model, and rely on the secular variation time scale $\tau = 415$ years, which we already used to scale time for this simulation. The idea is to test a filtering that mimics the sedimentary process and makes physical sense from the point of view of a 3D dynamo. For 3D dynamos, and for Earth's dynamo, the secular variation time scale defines the main time scale with which the non-dipole field is behaving (Lhuillier et al. 2011b). It provides a natural separation between the times scales of the long-term behavior of the dipole field, which is the one we are most interested in here, and its short time scales. Smoothing over a time period of 4τ typically removes such short time scales (Hulot & Le Mouël 1994). This corresponds to about 2 kyr. This is the value we tested, as it also is roughly consistent with the smoothing due to the sedimentary process in paleomagnetic reconstructions such as PADM2M and Sint-2000. For example, the regularization used to obtain the PADM2M reconstruction suppresses energy at time scales of 5-10 kyr (Ziegler et al. 2011). It finally is short enough compared to the decay time and event durations we identified for the field produced by the 3D (and DW) model (see Table 3.1). For consistency, we then also used the same time filtering to filter the time series produced by the DW model. In both cases, we used a moving average filter. Results are provided in Table 3.2, which lists the optimal MCC of threshold-based predictions for the DW and 3D models with and without smoothing for three prediction horizons. We found that the skill of threshold-based predictions only slightly increases for the 3D model, but hardly at all (to two digits) for the DW model. Thus, skills associated with the DW and 3D models are nearly unchanged by the smoothing process, and remain smaller than the skills associated with the P09 and G12 models.

Table 3.2: Maximum MCC of threshold-based predictions for the DW and 3D models with and without smoothing (smoothing window is $4\tau \approx 2$kyr) for three different prediction horizons. Optimal warning thresholds and MCC scores are computed over the entire run (no verification).

	Prediction horizon	0.5	1	1.5
DW	No smoothing	0.39	0.32	0.27
	2 kyr smoothing	0.39	0.32	0.27
3D	No smoothing	0.28	0.22	0.18
	2 kyr smoothing	0.32	0.24	0.19

3.4.4 Summary of results from the hierarchy of models

The hierarchy of models is consistent in that threshold-based predictions become more difficult, or, equivalently, less skillful, when the prediction horizon increases. This suggests that threshold-based predictions are at best useful for predicting low-dipole events with a lead time that is comparable to the average duration of the event (about 10 kyr on Earth's time scales). Moreover, the machinery of identifying thresholds by maximizing a skill score is robust in the sense that the skill during training is comparable to the skill during verification. Our overall approach is also robust with respect to the precise choices of start-of-event and end-of-event thresholds, and with respect to the choice of skill score (MCC, F_1 or CSI).

We observe strong differences in the skills of threshold-based predictions across the various models. The DW and 3D models exhibit complex behavior during reversals or excursions, with many polarity changes during the low-dipole event and the decay time is short compared to the event duration (fast reversals). The G12 model behaves differently: we do not observe quick polarity changes during a G12 reversal, no major excursions occur, and the decay time is larger than the event duration (slow reversals). The G12 model is more amenable to threshold-based predictions than the DW or 3D models, because of its simpler reversing behavior and because reversals are approached slowly. The P09 model falls in between the DW and 3D models and the G12 model.

Our numerical experiments with short training data sets, suggest that the main difficulty for

threshold-based predictions may not be the shortness of the observational record. The hierarchy of models is surprisingly consistent in that one may be able to determine useful warning thresholds, even if the training data are limited. The reasons for *why* this stability occurs, however, vary across the hierarchy of models. For the G12 model, low-dipole events are indeed easy to predict by a threshold and this threshold can be found by optimizing skill scores over short data sets. For the other models, the skill score is a nearly flat function of the threshold, i.e., different thresholds can lead to similar skill scores (recall Figure 3.11). More importantly, the overall skill of threshold-based predictions is low for the DW and 3D models, even when introducing some smoothing. Thus, threshold-based predictions may be of limited use for the DW and 3D models, because false positives and false negatives occur frequently. Again, the P09 model falls in between the G12 and DW/3D models.

We summarize our main results about threshold-based predictions for dipole models as follows.

(i) Across the hierarchy of models, the skill of threshold-based predictions degrades with the prediction horizon.

(ii) Across the hierarchy of models, threshold-based predictions are robust to minor variations of numerical details, such as choice of skill score (MCC or F1 or CSI), or choices of start-of-event and end-of-event thresholds.

(iii) Across the hierarchy of models, useful warning thresholds can be found even if the duration of the training period is short and comparable to the observational record. This suggests that the shortness of the observational record is not the main issue that makes computing warning thresholds difficult. The reasons for *why* this is the case, however, differ across the hierarchy of models.

(iv) The G12 model is more amenable (highest skill) to threshold-based predictions than the DW or 3D models (lowest skill). The skill of threshold-based predictions for the P09 model

falls in between the skills for G12 and DW/3D. Furthermore, we found that skills strongly correlate with the ratio of the average decay time to the average event duration.

3.5 Application to paleomagnetic reconstructions

We now take advantage of the lessons learned from the hierarchy of models and apply threshold-based predictions to the PADM2M and Sint-2000 paleomagnetic reconstructions, which provide proxies of the Earth's axial dipole intensity over the past 2 Myr (Ziegler et al. 2011; Valet et al. 2005). More specifically, PADM2M and Sint-2000 report the virtual axial dipole moment (VADM) in increments of 1 kyr for the past 2 Myr. We scale each reconstruction so that one unit of relative paleointensity corresponds to its time average ($5.32 \cdot 10^{22}$ Am^2 for PADM2M, $5.81 \cdot 10^{22}$ Am^2 for Sint-2000). The timing of reversals is based on the geomagnetic polarity time scale of Cande & Kent (1995), with a slight modification for the Cobb mountain sub-chron in the case of PADM2M (Morzfeld et al. 2017).

We note that PADM2M and Sint-2000 are "data" of the same process, namely Earth's dipole intensity over the past 2 Myr. Nonetheless, there are differences between PADM2M and Sint-2000, which are due to variations in the processing and interpretation of raw data, and also the raw data that goes into the two reconstructions. This means that differences between PADM2M and Sint-2000 indicate the level of uncertainty that is caused by difficulties with observing Earth's dipole over millions of years (see also Morzfeld & Buffett (2019)). Moreover, the fact that the observational record is short (2 Myr sampled in 1 kyr increments), implies that it is difficult to determine if any differences are (statistically) significant. It is important to keep this "minimum level of uncertainty" in mind when evaluating threshold-based predictions for the paleomagnetic reconstructions (note that we essentially treat the paleomagnetic reconstructions as "data," but we are aware that these reconstructions are themselves "models").

3.5.1 Event durations and decay times

Based on our definitions above, we compute the average and standard deviation of the event duration and decay times for the six events of PADM2M and Sint-2000. Results using ST = 10% and ET = 80% as before, are listed in Table 3.1 and these values should be compared with the corresponding values for the four models. In this context, it is important to realize that PADM2M and Sint-2000 never exhibit intensity values below 10% of their time average, which is why the definition of the low-dipole event in Section 3.3.1 contains the "or-statement": a low-dipole event starts when the intensity drops below the ST *or* if the dipole changes its sign.

We first note that PADM2M and Sint-2000 lead to results consistent with each other (e.g., average event duration and average decay times agree with each other within the corresponding standard deviations). We also note that both average event durations and average decay times fall within the range of values covered by the hierarchy of models. Hardly any model, however, leads to values satisfyingly matching those of PADM2M and Sint-2000 for both quantities. This is best seen in Figure 3.12, which shows the average decay time (ADT) plotted as a function of the average event duration (AED) for the four models and the paleomagnetic reconstructions.

The average event duration of PADM2M or Sint-2000 is longer than that of the G12 (shortest) and P09 models, and shorter than that of the DW and 3D (longest) models. We note, however, that associated standard deviations may reconcile the average event durations of the paleomagnetic reconstructions with those of the various models, but only marginally so for G12, which intrinsically displays little variation in the event duration. Moreover, the standard deviations for the event durations of the paleomagnetic reconstructions are quite comparable to those of the DW and 3D models, but larger than those of the P09 model, and are much larger than those of the G12 model. Overall, the average event duration of the paleomagnetic reconstructions lies in-between the average event durations of the G12/P09 and DW/3D models. We keep in mind that standard deviations for the paleomagnetic reconstructions may be corrupted by insufficient statistics, since the data document only six events.

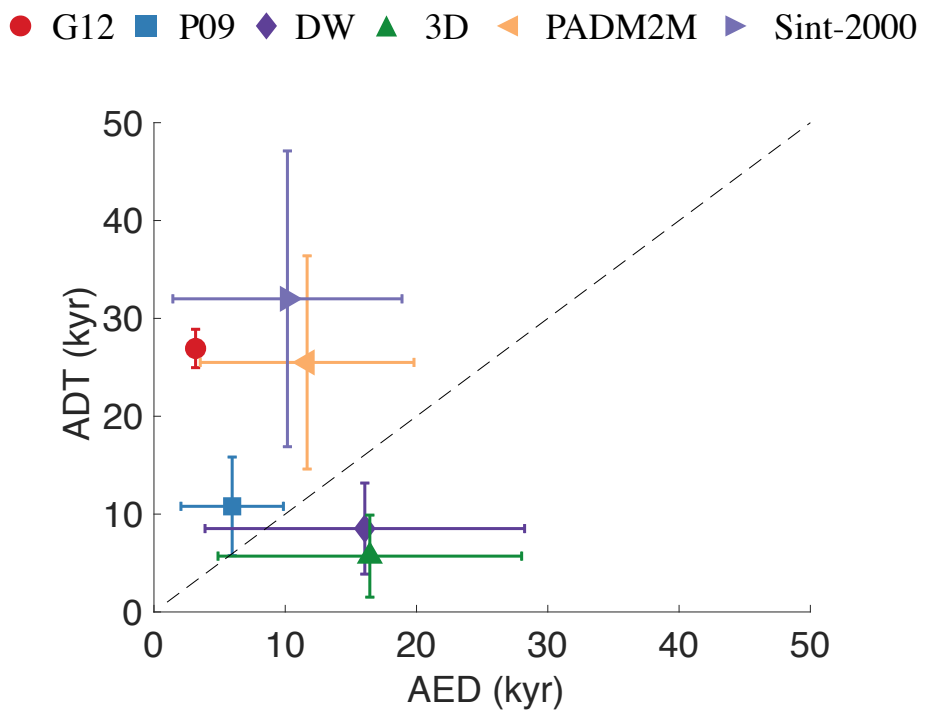

Figure 3.12: Average decay time (ADT) plotted as a function of the average event duration (AED) for the four models and the paleomagnetic reconstructions. Also shown are the error bars based on one standard deviation. In the case of the G12 model, the standard deviation of the event duration is too small to be visible as an error bar. Also shown is a 45° line that separates models or data for which ADT $>$ AED from models for which ADT $<$ AED.

The situation, however, is different when considering average decay times. The average decay times of the paleomagnetic reconstructions are much larger than those of the 3D (shortest), DW and P09 models, but comparable to that of the G12 model (longest). The standard deviations are much larger than that of the G12 model, and substantially larger than those of the P09, 3D and DW models (P09, DW and 3D models are comparable). This could be due to insufficient statistics or to data uncertainties, as suggested by the disagreement between the different values obtained with the PADM2M and Sint-2000 data sets. From the perspective of average decay times, it thus appears that the data are consistent with the G12 model.

Finally, we compute the ratio ρ of the average decay time to the average event duration for both paleomagnetic reconstructions. This leads to values of about two for PADM2M and three for Sint-2000 (see Table 3.1), which is consistent with the already known fact that intensity tends to decrease more slowly before a reversal than it recovers after it (Valet et al. 2005). The ratio ρ of PADM2M and Sint-2000 can also be compared to the corresponding ratios of the four models in Figure 3.9. We note that the ratios of the paleomagnetic reconstructions are much larger than the corresponding ratios associated with the 3D (smallest) and DW models; they are comparable to the corresponding ratio of the P09 model, and much smaller than the corresponding ratio of the G12 model.

3.5.2 Threshold-based predictions and their skills

We now apply threshold-based predictions to PADM2M and Sint-2000 using the same techniques as above and, as before, consider prediction horizons PH $= 0.5$, PH $= 1$ and PH $= 1.5$. Note that these PHs correspond to about 6 kyr, 11 kyr and 17 kyr in geophysical time. The ROC curves of threshold-based predictions for PADM2M and Sint-2000 are shown in Figures 3.13(a) and 3.13(b). These curves are computed over the entire 2 Myr time window covered by the paleomagnetic reconstructions. Inspecting the ROC curves qualitatively, we see that the skill of threshold-based predictions decreases with the prediction horizon. We observed this also for all

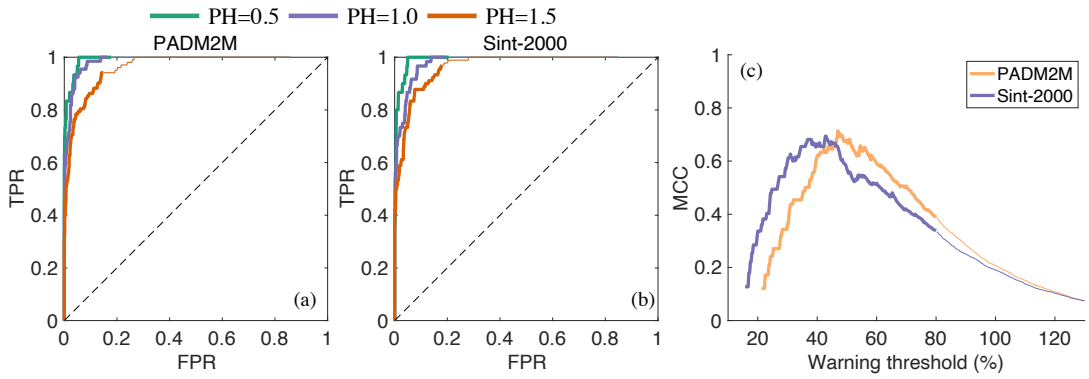

Figure 3.13: Panels (a) and (b): ROC curves for two paleomagnetic reconstructions and three prediction horizons, with PH = 0.5 (about 6 kyr) in green, PH = 1 (about 11 kyr) in purple, and PH = 1.5 (about 17 kyr) in orange. (a) PADM2M, (b) Sint-2000. An ROC curve is the collection of TPR/FPR pairs one obtains when varying the warning threshold. The thicker line corresponds to TPR/FPR pairs for which ST $<$ WT $<$ ET. The thin lines continue the ROC curves for ET $<=$ WT. Panel (c): MCC as a function of the warning threshold (prediction horizon is PH = 1). The ROC curves and MCC scores are computed over the entire 2 Myr covered by the paleomagnetic reconstructions.

four models. Comparing the ROC curves of the paleomagnetic reconstructions in Figures 3.13(a) and (b) with the ROC curves of the models in Figure 3.6, the ROC curves of the paleomagnetic reconstructions resemble those of the P09 model. Figures 3.13(c) shows the curve traced out by the MCC as when varying the warning threshold for PADM2M and Sint-2000 (MCC is computed over the entire 2 Myr time window). Again, we note that the curves corresponding to the paleomagnetic reconstructions are qualitatively similar to the corresponding curve of the P09 model (see Figure 3.11).

We also compute optimal MCC scores for PADM2M and Sint-2000 via training and verification. We use the first 0.95 Myr, containing four events, for training (finding an optimal warning threshold) and use the remaining 1.05 Myr, containing two events, for verification. Table 3.1 lists these MCCs for PADM2M and Sint-2000, together with the MCCs of the four models, when computed with training data containing a comparable number of events (four events during training for the paleomagnetic data and five events during training for the models, see Section 3.4.2). Figure 3.14 shows the (verification) MCC scores for PADM2M and Sint-2000

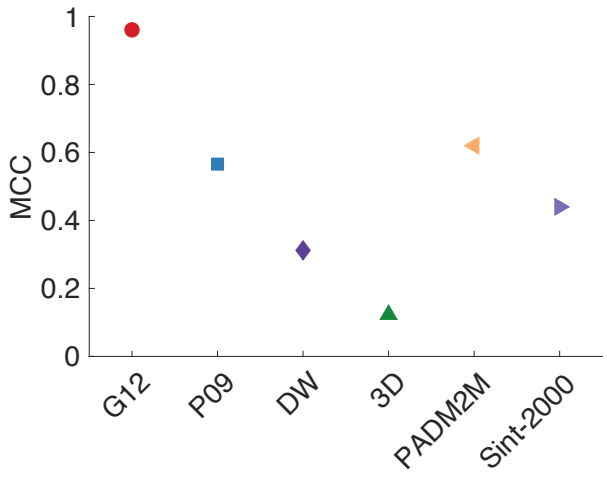

Figure 3.14: MCC of four models and two paleomagnetic reconstructions (PADM2M and Sint-2000). The optimal WT is computed using training data containing five events in the case of the models, and four events in the case of the paleomagnetic reconstructions (see text for details).

along with those of the models.

We first note from Table 3.1 that the MCC skill score drops from training to verification and that the verification skills for PADM2M and Sint-2000 are quite different. This is caused by the verification periods being extremely short, with only two events during verification. Thus, while the warning threshold we find from a limited observational record may be quite accurate, it remains difficult to evaluate the skill of threshold-based predictions. These difficulties are due to the shortness of the observational record – we only have 2 Myr, with six events, to base our training *and* validation on. Nevertheless, we again find that paleomagnetic reconstructions tend to produce verification MCC scores quite consistent with what could be anticipated based on our analysis of the ratio of the average decay time to the average event duration. The MCC associated with the paleomagnetic reconstructions, indicative of the skill of intensity threshold based prediction, is indeed larger than the MCC recovered for the 3D (smallest) and DW models, comparable to that of the P09 model, and much smaller than that of the G12 model.

We illustrate threshold-based predictions for the paleomagnetic reconstructions and PH $=$ 1 in Figure 3.15. This figure first confirms that the choice of the start-of-event (ST) and end-of-

event (ET) thresholds properly identifies the six events of interest. There are five reversals and one major excursion, which corresponds to what is known as the Cobb mountain subchron at 1.19 Myr, and is indeed an event during which the field temporarily changed its polarity at low intensity. This figure also illustrates the limitations of threshold-based predictions when using PADM2M or Sint-2000. In the case of PADM2M, with an optimal warning threshold of $\hat{\mathrm{WT}}_{\mathrm{PADM2M}} = 50.75\%$ (corresponding to $2.70 \cdot 10^{22}$ Am^2), we note the occurrence of two instances of false positives, where no low-dipole event is observed, but a low-dipole event is predicted, (near the -1.5 Myr and -0.25 Myr marks). One of these instances of false positives occurs during training, the other during verification. Such false positives do not occur in the case of the Sint-2000, which also has a lower optimal WT of $\hat{\mathrm{WT}}_{\text{Sint-2000}} = 36.75\%$ (corresponding to $2.14 \cdot 10^{22}$ Am^2). One may thus intuitively expect that the predictions will have a lower skill when applied to PADM2M than to Sint-2000, but in fact this is not the case: the skill during verifications is higher for PADM2M than for Sint-2000, but the skill for training is higher for Sint-2000 than for PADM2M. This is perhaps counter intuitive because one is tempted to think of false positives that occur "far" from a reversal as more severe than false positives or false negatives that occur "close" to a reversal. The MCC score, however, does not apply special meaning to the categories of "positive" and "negative," so that predictions of the timing of the two reversals, e.g., during verification, are more accurate for PADM2M than for Sint-2000.

The ROC curves and the MCC skill scores for the paleomagnetic reconstructions and models suggest that the predictive skill of threshold-based predictions of the paleomagnetic reconstructions may be comparable to the skill of these predictions for the P09 model. Because it is difficult to verify threshold-based predictions using the observational record only, we may use the P09 model to investigate the skill of threshold-based predictions, applied to the paleomagnetic reconstructions. Threshold-based predictions (PH = 1) for the P09 model are illustrated in Figure 3.7 (note that the predictions in Figure 3.7 make use of a large training data set). Indeed, when training threshold-based predictions for P09 with training data that contains five low-dipole

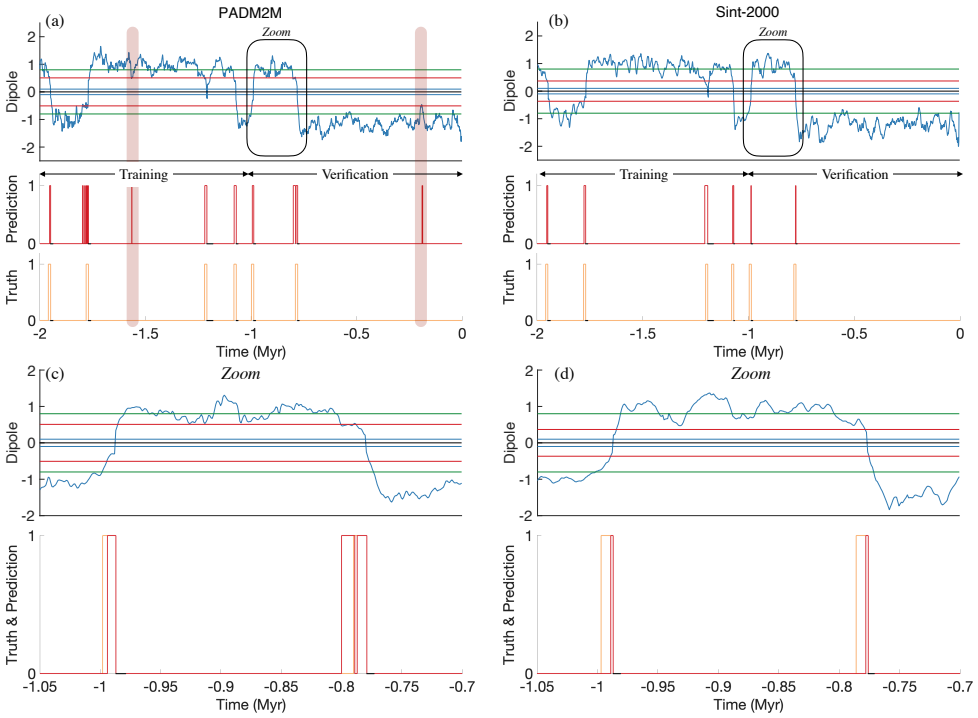

Figure 3.15: Illustration of threshold-based predictions for PADM2M ((a) and (c)) and Sint-2000 ((b) and (d)) reconstructions. The prediction horizon is PH = 1 (about 11 kyr) and the optimal warning threshold computed over 0.95 Myr of training data, containing four events. The corresponding warning thresholds (expressed in percent of the average intensity) are $\hat{\text{WT}}_{\text{PADM2M}} = 50.75\%$ ($2.70 \cdot 10^{22}$ Am^2) and $\hat{\text{WT}}_{\text{Sint-2000}} = 36.75\%$ ($2.14 \cdot 10^{22}$ Am^2) for respectively PADM2M and Sint-2000. Panels (a) and (b) contain three subfigures. *Top*. Blue: dipole time series. Blue/green/red horizontal lines: start-of-event/end-of-event/warning thresholds. *Center*. Graphs are zero if the threshold-based prediction is "no low-dipole event will start during the prediction horizon;" graphs are one if the threshold-based prediction is "a low-dipole event will start during the prediction horizon." *Bottom*. Graphs are zero if no low-dipole event starts within the prediction horizon; graphs are one if a low-dipole event starts during the prediction horizon. Panels (c) and (d) show magnifications during a time interval that includes the two reversals that occur during verification.

events (comparable to paleomagnetic reconstructions), the optimal WT of the P09 model of 54.5% is quite comparable to that obtained for PADM2M (50.75%) and slightly more than that obtained for Sint-2000 (36.75%), all of which are consistent with the range of values found in Figures 3.10 and 3.11. We also observe that the predictions for P09 are similar to the predictions for the paleomagnetic reconstructions. We encounter a large number of true negatives, several false negatives, for which the threshold-based predictions trigger a little too late, and occasionally encounter false positives that occur during periods when no low-dipole event occurs.

In summary, we conclude that threshold-based predictions are feasible for the paleomagnetic reconstructions, but lead to moderate success. They share similar characteristics as threshold-based predictions for the P09 model, and suffer from similar caveats:

(i) Low-dipole events can be predicted only a relatively short time ahead, i.e., the prediction horizon should be about one average event duration or less. On Earth's time scale, this means the prediction horizon should be about 10 kyr or less.

(ii) Low-dipole events may be predicted a few kyr too late (false negatives), which is significant in view of the relatively short prediction horizon.

(iii) One must be prepared for false positives to occur even when no low-dipole event is about to happen.

The above conclusions are supported by two paleomagnetic reconstructions, PADM2M and Sint-2000, but threshold-based predictions show some sensitivity to which reconstruction we use. This is perhaps best illustrated by the predictions in Figure 3.15, but it is also clear from the skill scores in Table 3.1. As indicated above, differences between results stemming from PADM2M or Sint-2000 establish an uncertainty that cannot be resolved, because this uncertainty is caused by our limited ability to observe Earth's dipole over millions of years. In this context, we wish to point out that we did not use other global models, e.g., PISO-1500 (Channell et al. 2009), because it is biased towards the North Atlantic region due to the fact that only stacks

with a high sedimentation rate are used (see, e.g., Figure 5 of Panovska et al. (2019)). Indeed, PISO-1500 is less representative of the (global) axial dipole field than PADM2M and Sint-2000 (see also Ziegler et al. (2011)). Exploring the consequences of such differences is beyond the scope of our work.

Finally, we want to bring a few details to the reader's attention. In particular, we want to emphasize that the average dipole intensity and the average event duration for threshold-based predictions for PADM2M or Sint-2000 are computed using the entire 2 Myr record. One could also envision to compute the average intensity based on training data only. We decided not to do so for the following reasons. The average intensity defines the start and end of an event, because start-of-event and end-of-event thresholds are defined in terms of the average intensity. The average event duration, and even the number of events, are implicitly defined by the start-of-event and end-of-event thresholds and, thus, also depends on the average intensity. The average event duration, in turn, is used in the definition of the prediction horizon. In summary, the average intensity directly affects (i) the number of events; (ii) the average event duration; and (iii) the prediction horizon. By computing the average intensity over the 2 Myr reconstructions, we have assumed these difficulties away and fix the average intensity a priori. We find that this is more practically relevant because the average intensity may be determined by using additional information. Nonetheless, we also made threshold-based predictions for which we compute the average event duration based on training data and the results are nearly identical to the results we show above.

3.6 Concluding comments

The main purpose of this study is to test the possibility that a low value of the axial dipole intensity could be used as a natural indicator of an upcoming dipole reversal. To answer this question, we analyzed a hierarchy of numerical models, and Earth's axial dipole field as

documented by the PADM2M and Sint-2000 paleomagnetic VADM reconstructions (Ziegler et al. 2011; Valet et al. 2005). More specifically, we test the possibility of relying on an intensity threshold-based strategy, whereby once the axial dipole intensity drops below a warning threshold, it is predicted that the intensity will drop further and lead to a low-dipole event (either a reversal or a major excursion) within some specified time, called the prediction horizon. Although the principle of such a strategy appears to be fairly intuitive, implementing it in a robust way led us to introduce a dedicated methodology.

Our method requires that we define a warning threshold (WT), a start-of-event threshold (ST), an end-of-event threshold (ET) and a prediction horizon (PH). Both ST and ET appear to be most conveniently defined in terms the average intensity of the axial dipole (in practice ST=10% and ET=80%). ST and ET also define an average event duration (AED, average time elapsed between when the intensity passes below the ST and when it recovers back to above the ET). The prediction horizon is defined in terms of the average event duration and we consider predictions with PHs of about one AED. Having chosen the ST, ET and PH, we identify the warning threshold (WT) by maximizing a skill score. Several skill scores have been tested, and all adequate choices led to similar conclusions. Similarly, we showed that the exact choices of the ST and ET percentages are not critical, provided these properly bracket the events of interest. The code we use to implement the prediction is available on github https://github.com/kjg136/Threshold. We archived the code used to generate all figures in https://doi.org/10.5281/zenodo.4267116.

A first major conclusion is that the skills of intensity threshold-based predictions vary surprisingly widely within the hierarchy of numerical models we investigated (G12, P09, DW and 3D models). The only model that leads to a high skill (implying that the intensity threshold-based predictions are reliable) is the G12 model. This result is in line with the results obtained by Morzfeld et al. (2017), who identified a high skill of intensity threshold-based predictions for this model, using a simpler strategy and a less robust analysis. All other models lead to lower skills, implying that the intensity threshold-based predictions are less reliable. This is, again, consistent

with Morzfeld et al. (2017), who investigated the P09 model and a model (B13, Buffett et al. (2013)) similar to the DW model, but did not investigate the 3D model. In the present study, we were able to rank these skills more accurately and identify one key property that may play a major role in defining the skills of threshold-based predictions in the context of numerical dynamos and VADM reconstructions (PADM2M and Sint-2000).

This key property is that skills of intensity threshold-based predictions correlate with the ratio of the average decay time (defined as the time between the start of the event and the most recent time instance at which the intensity is equal to the end-of-event threshold) to the average event duration. The larger this ratio, the better the skill. The models and the PADM2M and Sint-2000 reconstructions are consistent with this rule. As already noted, this asymmetry between the way the field decreases towards a reversal and the way it recovers its strength after the reversal is a well-known property of the field (Valet et al. 2005), which in fact may be related to the more general tendency of the Earth's magnetic field to spend more time decreasing than increasing at any time (see, e.g., Ziegler & Constable 2011; Avery et al. 2017). What the present study thus suggests is that this slight asymmetry is what defines the skill of intensity threshold-based predictions when applied to Earth's magnetic field. Unfortunately, because this ratio is about two to three, the skill of threshold-based predictions is limited. As our study further shows, this, more than the relatively short duration of the Sint-2000 and PADMD2M reconstructions, is what likely makes intensity threshold-based predictions using these data modestly reliable.

Despite the limitations we identified for intensity threshold-based predictions, it is worth pointing out that today's axial dipole field, with a magnitude of about $7.8 \cdot 10^{22} \mathrm{Am}^2$ (Constable & Korte 2006), is much larger than the warning thresholds we identified by using either Sint-2000 ($\hat{\mathrm{WT}}_{\text{Sint-2000}} = 36.75\%$ of the average $5.81 \cdot 10^{22}$ Am^2, which amounts to $2.14 \cdot 10^{22}$ Am^2) or PADM2M ($\hat{\mathrm{WT}}_{\text{PADM2M}} = 50.75\%$ of the average $5.32 \cdot 10^{22}$ Am^2, amounting to $2.70 \cdot 10^{22}$ Am^2). Intensity threshold-based predictions thus suggest that no low-dipole event will occur within the next 10 kyr. This is in line with many other recent predictions, see, e.g., Constable & Korte

(2006); Morzfeld et al. (2017); Brown et al. (2018).

As an interesting additional outcome of this study, we note that testing the skills of threshold-based predictions on numerical dynamos is a fairly discriminating way of testing the Earth-like nature of the axial dipole field behavior of the models. This skill is distinct from the ability of numerical simulations to reproduce the frequency with which reversal occurs. This is evident from the fact that threshold-based predictions have different skills for the DW and P09 models, whereas both models are characterized by reversal frequencies comparable to that of the Earth over the last 25 My (about 5 reversals per Myr). As this skill appears to be correlated with the ratio of the average decay time to the average event duration (a measure of the asymmetry with which the field evolves towards a reversal and next recovers its full strength), it also appears to be distinct from other criteria often used to characterize the Earth's dipole field behavior, such as its frequency content (Constable & Johnson 2005), or the relative time spent in transitional periods (based on dipole latitudes being less than 45°), as recently suggested by Sprain et al. (2019). Furthermore, in spite of its favorable ratings according to the criteria defined by Christensen et al. (2010) for the recent field, and Sprain et al. (2019) for the paleomagnetic field (recall section 3.2.2), the field produced by the 3D model appears to not match that of the Earth's field (as described by PADM2M and Sint-2000) in terms of intensity threshold-based prediction skill (and ratio of the average decay time to the average event duration). In agreement with the suggestions of Ziegler & Constable (2011) and Avery et al. (2017), and since it appears to play a significant role in the way reversals occur, we strongly encourage the community to also consider predictive skills and asymmetric temporal behavior as additional criteria to identify Earth-like dynamo simulations.

The present study shows that intensity threshold-based predictions of reversals appear to be of only limited value, but we emphasize that we investigated these limitations for only one specific threshold-based prediction, namely predicting whether a reversal or major excursion occurs during a specified time window. Other types of predictions might deal instead with the

probability of a reversal or major excursion during a specified time window. In this case, a large number of reversals would be needed to test these predictions. It is also worthwhile to comment on other routes to more robust and reliable predictions. Taking advantage of machine-learning and deep learning could be a possibility (Goodfellow et al. 2016). In this context, however, one should be careful to check that the shortness of the paleomagnetic reconstructions is not a limiting factor, as deep learning is known to work best when data availability is vast, and only poorly when data are limited. Another approach is to rely on merging the observations in a process called data assimilation (DA) (see, e.g., Carrassi et al. 2018). This strategy has been successful in numerical weather prediction because the atmospheric model is of high quality, and because observations of the atmospheric state are plentiful (Bauer et al. 2015). It currently is developing in the field of geomagnetism (Fournier et al. 2010). Using DA for predicting dipole reversals, however, is difficult due to the lack of a suitable 3D model that can be run fast enough and the fact that the observations are limited to the virtual axial dipole moment over 2 Myr. Here, the main difficulty lies in identifying, or creating, useful models that are simple enough to allow for data assimilation but complex enough to represent all relevant time scales. Nevertheless, Morzfeld et al. (2017) recently showed that using such an approach with the G12 model and assimilating either PADM2M or Sint-2000, could lead to some success. No similar success could be reached with the P09 model, which was also tested. In that approach, indeed, the key to success appears to be the dynamical way the axial dipole produced by the model approaches reversals. It appears that the way the G12 model approaches reversals is more similar to how Earth's axial dipole field approaches reversals, than the P09 model. This leads to the interesting possibility of finding a better suited low-dimensional model with properties intermediate between the G12 model (whose decay-time properties make it well suited for data assimilation) and P09 (with intensity threshold-based prediction properties closest to that of the paleomagnetic reconstructions) leading to better predictions of reversals several kyr ahead.

Acknowledgements

KG acknowledges that this work was supported by NASA Headquarters under the NASA Earth and Space Science Fellowship Program - Grant "80NSSC18K1351". This work was supported in part by the French Agence Nationale de la Recherche under grant ANR-19-CE31-0019 (revEarth). All authors would like to thank Nathanael Schaeffer (ISTerre, CNRS, Université Grenoble Alpes) and Thomas Gastine (Université de Paris, Institut de Physique du Globe de Paris) for allowing us to use the dipole time series of the 3D model. We acknowledge GENCI for access to the Irene resource (TGCC) under grants "Grand Challenge" GCH0315 and A0060407382. We thank Maggie S. Avery (UC Berkeley) and an anonymous reviewer for helping us improve this paper. We thank Cathy Constable (Scripps Institution of Oceanography, University of California, San Diego) and Bruce Buffett (UC Berkeley) for meaningful discussions. AF thanks Richard Bono and Courtney Sprain for their assistance in the calculation of ΔQ_{PM}. All authors contributed to the ideas behind the approach taken in the manuscript and all authors contributed to writing the paper; KG wrote the code.

Chapter 3, in full, is a reprint of material as it appears in Gwirtz, K., Morzfeld, M., Fournier, A., and Hulot, G., Can one use Earth's magnetic axial dipole field intensity to predict reversals?, *Geophysical Journal International*, **225** (1), 277-297, (2021)). The dissertation author was the primary investigator and author of this paper.

Appendix

Table 3.3: Acronyms used in this paper

Type	Acronym	Explanation
Outcomes of events	P	Number of positives
	N	Number of negatives
Outcomes of predictions	TP	True positive
	FP	False positive
	TN	True negative
	FN	False negative
Receiver operator charecteristics	TPR	True positive rate (equation (3.9))
	FPR	False positive rate (equation (3.9))
	ROC	Receiver operator characteristic
Skill scores	ACC	Accuracy (equation (3.5))
	F_1	F_1 skill score (equation (3.6))
	CSI	Critical success index (equation (3.7))
	MCC	Mathews correlation coefficient (equation (3.8))
Threshold-based predictions	ST	Start-of-event threshold
	ET	End-of-event threshold
	WT	Warning threshold
	PH	Prediction horizon
	AED	Average event duration
	ADT	Average decay time
	$\rho = \frac{\text{ADT}}{\text{AED}}$	ratio of ADT and AED
Models	G12	Differential equation model (Gissinger 2012)
	P09	Stochastic model (Pétrélis et al. 2009)
	DW	Stochastic double well model (Morzfeld & Buffett 2019)
	3D model	3-dimensional dynamo simulation (unpublished)
	SDE	Stochastic differential equation
	MHD	Magneto-hydrodynamic
	DA	Data assimilation
Data	VADM	Virtual axial dipole moment
	PADM2M	VADM reconstruction (Ziegler et al. 2011)
	Sint-2000	VADM reconstruction (Valet et al. 2005)

References

Aubert, J., 2019. Approaching Earth's core conditions in high-resolution geodynamo simulations, *Geophys. J. Int.*, **219**, S137–S151.

Avery, M. S., Gee, J. S., & Constable, C. G., 2017. Asymmetry in growth and decay of the geomagnetic dipole revealed in seafloor magnetization, *Earth planet. Sci. Lett.*, **467**, 79 – 88.

Barrett, H. & Myers, K., 2003. *Foundations of Image Science*, Wiley.

Bauer, P., Thorpe, A., & Brunet, G., 2015. The quiet revolution of numerical weather prediction, *Nature*, **525**(7567), 47.

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.

Buffett, B., 2015. Dipole fluctuations and the duration of geomagnetic polarity transitions, *Geophys. Res. Lett.*, **42**, 7444–7451.

Buffett, B. & Matsui, H., 2015. A power spectrum for the geomagnetic dipole moment, *Earth planet. Sci. Lett.*, **411**, 20–26.

Buffett, B., Ziegler, L., & Constable, C., 2013. A stochastic model for paleomagnetic field variations, *Geophys. J. Int.*, **195**(1), 86–97.

Buffett, B. A., King, E. M., & Matsui, H., 2014. A physical interpretation of stochastic models for fluctuations in the Earth's dipole field, *Geophys. J. Int.*, **198**(1), 597–608.

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

Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G., 2018. Data assimilation in the geosciences: An overview of methods, issues, and perspectives, *Wiley Interdisciplinary Reviews: Climate Change*, **9**(5), e535.

Channell, J., Xuan, C., & Hodell, D., 2009. Stacking paleointensity and oxygen isotope data for the last 1.5Myr (PISO-1500), *Earth Planet. Sci. Lett*, **283**(1), 14 – 23.

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

Chorin, A. & Hald, O., 2013. *Stochastic tools in mathematics and science*, Springer, 3rd edn.

Christensen, U. R. & Wicht, J., 2015. Numerical dynamo simulations, in *Core Dynamics*, vol. 8 of **Treatise on Geophysics**, chap. 8, pp. 245–277, eds Olson, P. & Schubert, G., Elsevier, 2nd

edn.

Christensen, U. R., Aubert, J., & Hulot, G., 2010. Conditions for Earth-like geodynamo models, *Earth planet. Sci. Lett.*, **296**(3-4), 487–496.

Constable, C. & Johnson, C., 2005. A paleomagnetic power spectrum, *Phys. Earth planet. Inter.*, **153**, 61–73.

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

Fawcett, T., 2006. An introduction to ROC analysis, *Pattern Recognit. Lett.*, **27**(8), 861 – 874, ROC Analysis in Pattern Recognition.

Finlay, C. C., Aubert, J., & Gillet, N., 2016. Gyre-driven decay of the Earth's magnetic dipole, *Nature Communications*, **7**, 10422.

Fournier, A., Hulot, G., Jault, D., Kuang, W., Tangborn, W., Gillet, N., Canet, E., Aubert, J., & Lhuillier, F., 2010. An introduction to data assimilation and predictability in geomagnetism, *Space Sci. Rev.*, **155**, 247–291.

Gissinger, C., 2012. A new deterministic model for chaotic reversals, *Eur. Phys. J. B.*, **85**, 137.

Glatzmaier, G. & Coe, R., 2015. Magnetic polarity reversals in the core, in *Core Dynamics*, vol. 8 of **Treatise on Geophysics**, chap. 9, pp. 279–295, eds Olson, P. & Schubert, G., Elsevier, 2nd edn.

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

Hoyng, P., Ossendrijver, M., & Schmitt, D., 2001. The geodynamo as a bistable oscillator, *Geophys. Astrophys. Fluid Dyn.*, **94**, 263–314.

Hulot, G. & Le Mouël, J.-L., 1994. A statistical approach to the Earth's main magnetic field, *Phys. Earth planet. Inter.*, **82**, 167–183.

Hulot, G., Eymin, C., Langlais, B., Mandea, M., & Olsen, N., 2002. Small-scale structure of the Geodynamo inferred from Oersted and Magsat satellite data, *Nature*, **416**, 620–623.

Hulot, G., Finlay, C. C., Constable, C. G., Olsen, N., & Mandea, M., 2010a. The magnetic field of planet Earth, *Space Sci. Rev.*, **152**, 159–222.

Hulot, G., Lhuillier, F., & Aubert, J., 2010b. Earth's dynamo limit of predictability, *Geophys. Res. Lett.*, **37**, L06305.

Joliffe, I., 2016. The dice co-efficient: a neglected verification performance measure for deterministic forecasts of binary events, *Meteorol. Appl.*, **23**, 89–90.

Kenney, J. & Keeping, E., 1966. *Mathematics of Statistics, Pt. 1*, Van Nostrand Company, 3rd edn.

Kloeden, P. E. & Platen, E., 1999. *Numerical solution of stochastic differential equations*, Springer.

Laj, C. & Kissel, C., 2015. An impending geomagnetic transition? Hints from the past, *Front. Earth Sci.*, **3**, 61.

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**, L09306.

Lhuillier, F., Hulot, G., & Gallet, Y., 2013. Statistical properties of reversals and chrons in numerical dynamos and implications for the geodynamo, *Phys. Earth planet. Inter.*, **220**, 19–36.

Lowrie, W. & Kent, D., 2004. Geomagnetic polarity time scale and reversal frequency regimes, *Timescales of the paleomagnetic field*, **145**, 117–129.

Meduri, D. & Wicht, J., 2016. A simple stochastic model for dipole moment fluctuations in numerical dynamo simulations, *Front. Earth Sci.*, **4**, 38.

Morzfeld, M. & Buffett, B. A., 2019. A comprehensive model for the kyr and Myr timescales of Earth's axial magnetic dipole field, *Nonlinear Proc. Geoph.*, **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.

Ogg, J., 2012. Geomagnetic polarity time scale, in *The geologic time scale 2012*, chap. 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.

Olson, P., Deguen, R., Hinnov, L. A., & Zhong, S., 2013. Controls on geomagnetic reversals and core evolution by mantle convection in the phanerozoic, *Phys. Earth planet. Inter.*, **214**, 87–103.

Panovska, S., Korte, M., & Constable, C., 2019. One hundred thousand years of geomagnetic

field evolution, *Rev. Geophys.*, **57**(4), 1289–1337.

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

Schaeffer, N., 2013. Efficient spherical harmonic transforms aimed at pseudospectral numerical simulations, *Geochemistry, Geophysics, Geosystems*, **14**(3), 751–758.

Schaeffer, N., Jault, D., Nataf, H.-C., & Fournier, A., 2017. Turbulent geodynamo simulations: a leap towards Earth's core, *Geophys. J. Int.*, **211**(1), 1–29.

Schmitt, D., Ossendrijver, M., & Hoyng, P., 2001. Magnetic field reversals and secular variation in a bistable geodynamo model, *Phys. Earth planet. Inter.*, **125**, 119–124.

Sprain, C. J., Biggin, A. J., Davies, C. J., Bono, R. K., & Meduri, D. G., 2019. An assessment of long duration geodynamo simulations using new paleomagnetic modeling criteria (Q_{PM}), *Earth planet. Sci. Lett.*, **526**, 115758.

Valet, J.-P. & Fournier, A., 2016. Deciphering records of geomagnetic reversals, *Rev. Geophys.*, **54**(2), 410–446, 2015RG000506.

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

Valet, J.-P., Fournier, A., Courtillot, V., & Herrero-Bervera, E., 2012. Dynamical similarity of geomagnetic field reversals, *Nature*, **490**, 89–93.

Wicht, J. & Sanchez, S., 2019. Advances in geodynamo modelling, *Geophys. Astrophys. Fluid Dyn.*, **113**(1-2), 2–50.

Ziegler, L. & Constable, C., 2011. Asymmetry in growth and decay of the geomagnetic dipole, *Earth planet. Sci. Lett.*, **312**(3), 300–304.

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.

Chapter 4

Can machine learning find precursors to reversals of Earth's magnetic dipole field?

K. Gwirtz*, T. Davis* and M. Morzfeld*
* Cecil H. and Ida M. Green Institute of Geophysics and Planetary Physics,
Scripps Institution of Oceanography, University of California, San Diego;

Abstract. We use machine learning to search for precursors to reversals of Earth's magnetic dipole field, in time series of dipole intensity. To do this, machine learning models are trained to classify segments of time series of dipole intensity according to whether they precede a reversal. We evaluate if the machine learning models have successfully identified precursors to reversals by testing their predictive skill on independent validation data—if precursors have been found, predictive skill should be high. Given the limited observational record, we first test machine learning predictions with simulations from a simplified model of the dipole and a numerical 3-D geodynamo, before studying paleomagnetic reconstructions of the Virtual Axial Dipole Moment (PADM2M and Sint-2000). It is found that predictive skill of machine learning varies among the dipole models and reconstructions. We present evidence that this not the result of the particular machine learning methods chosen but simply reflects the differing extent to which precursors to oncoming reversals exist in the simulations and reconstructions. We conclude that machine learning may be of limited use, particularly in view of the small amount of data available for training and validation.

4.1 Introduction

We wish to predict reversals in the polarity of the Earth's magnetic dipole. Such reversals have occurred numerous times throughout Earth's history (Ogg 2012; Cande & Kent 1995; Lowrie & Kent 2004), most recently around 780 kyr ago. Predictions of reversals by a threshold strategy were recently studied in Gwirtz et al. (2021). In simple terms, the threshold prediction strategy is: when the dipole intensity drops below a prescribed threshold, predict a reversal (see section 4.2). We extend this work by studying predictions which are informed not only by the present dipole intensity, but its recent history. To do this, we use machine learning techniques to search for precursors to reversals in time series of axial dipole intensity. We primarily employ support-vector machines (SVMs, see, e.g., Cristianini et al. 2000) but also use long short-term memory networks (LSTMs, see, e.g., Hochreiter & Schmidhuber 1997) to evaluate the robustness of results.

The effectiveness of machine learning techniques can depend on the availability of large data sets for training. However, the observational record of Earth's magnetic axial dipole field is limited—the paleomagnetic reconstructions we consider cover just the last 2 Myr (PADM2M and Sint-2000). For this reason, we first study machine learning predictions of reversals using numerical simulations because these "data" are not limited. Representing the physics of the Earth's main magnetic field in a model calls for a 3-D numerical dynamo capable of simulating the turbulent flow in the outer core and its coupling to the magnetic field. Unfortunately, numerical dynamos are computationally demanding and the number of simulations which produce reversals is small (Lhuillier et al. 2013; Olson et al. 2013). Alternatively, a number of low-dimensional models exist which aim to capture the behavior of Earth's magnetic dipole, including reversals, at a much lower computational cost (Hoyng et al. 2001; Schmitt et al. 2001; Pétrélis et al. 2009; Gissinger 2012; Buffett et al. 2013, 2014; Buffett & Matsui 2015; Buffett 2015; Meduri & Wicht 2016; Morzfeld & Buffett 2019). To study reversal predictions without the restrictions of limited data, in addition to paleomagnetic reconstructions, we use simulations from a reversing 3-D

dynamo and a low-dimensional model.

For both the observational record and numerical simulations, the rarity of reversals creates an additional difficulty in the application of machine learning. We pose the challenge of predicting reversals as a classification problem (Goodfellow et al. 2016). In simple terms, we train machine learning models to classify segments of a dipole intensity time series as either positive (P), a reversal occurs in the near future, or negative (N), a reversal does not occur in the near future. Because reversals are rare, a majority of segments taken from a time series of dipole intensity will belong to the class of N. In this framework, the data is said to be *imbalanced*. The difficulty of training standard machine learning models with imbalanced data is that they tend to favor strategies which maximize accuracy and therefore, may learn to always assign input to one class. For example, in the prediction of reversals, one can always achieve a high degree of accuracy by predicting no reversal is to occur in the near future (Gwirtz et al. 2021). To address this issue we penalize false negatives (failing to correctly predict a reversal) more severely than false positives (incorrectly predicting a reversal will occur) during the training of machine learning models.

The remainder of this paper is organized in the following way. In section 4.2 we introduce the models and paleomagnetic reconstructions that we use, provide relevant background from the earlier reversal prediction study of Gwirtz et al. (2021), and briefly outline SVMs and LSTMs. In section 4.3 we detail the procedures we use to train and validate machine learning models. The results of a collection of numerical experiments are then presented in section 4.4 followed by a summary and discussion of the study in section 4.5.

4.2 Background: Models, paleomagnetic reconstructions, threshold-based predictions and machine learning

We provide the background needed to understand the numerical experiments of the paper, beginning with a brief overview of the numerical models and paleomagnetic reconstructions

that we use. Next we present a review of previous work concerning threshold-based predictions, on which we build. Specifically, we provide the definitions of *low-dipole events* and *prediction horizons* which we use, and describe the mechanics of the threshold strategy which we compare machine learning predictions with. This is followed by an outline of the machine learning methods we employ.

4.2.1 Numerical models and paleomagnetic reconstructions

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

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \mu Q - VD, \quad \frac{\mathrm{d}D}{\mathrm{d}t} = -\nu D + VQ, \quad \frac{\mathrm{d}V}{\mathrm{d}t} = \Gamma - V + QD, \tag{4.1}$$

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 with a change in sign corresponding to a reversal. We refer to this as the G12 model and use the G12 millenium timescale (1 dimensionless time unit = 4 kyr) of Morzfeld et al. (2017). A segment of the dipole intensity of a G12 simulation is shown in the top panel of figure 4.1. Numerical simulations are performed via a fourth-order Runge-Kutta scheme.

The other simulation we consider is from a 3-D numerical dynamo which exhibits polarity reversals. The time series of the axial dipole from this simulation was previously used in Gwirtz et al. (2021) where further details of the numerical model and its favorable 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). A segment of the time series of axial dipole intensity from this simulation can be seen in the bottom panel of figure 4.1.

In addition to the numerical G12 and 3-D models, we also use the PADM2M and Sint-2000 paleomagnetic reconstructions (Ziegler et al. 2011; Valet et al. 2005) which consist of estimates

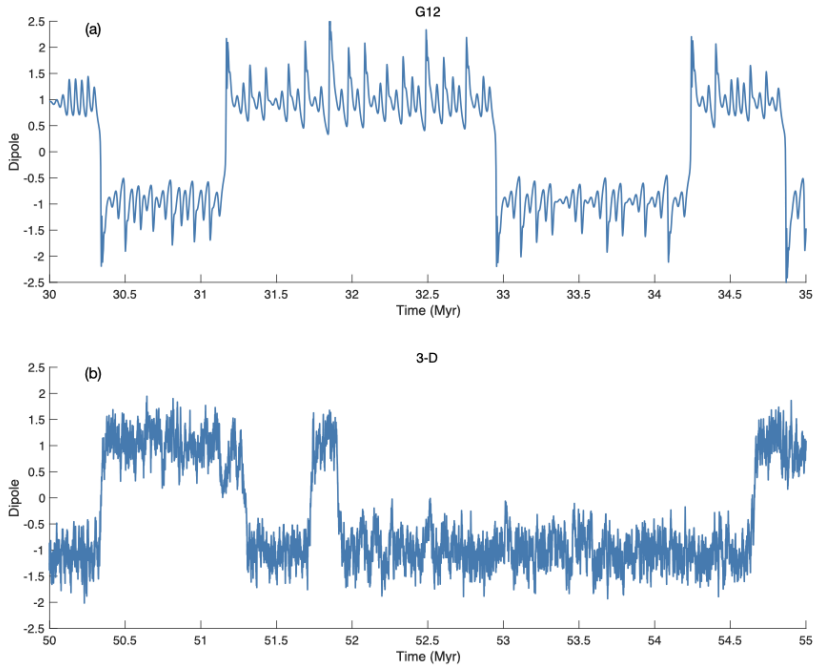

Figure 4.1: Dipole intensity as a function of time for the (a) G12 and (b) 3-D models. The sign indicates polarity and the amplitude is scaled such that the average intensity is one.

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 have a time step of 1kyr and can be seen in figure 4.2. We note that despite covering the same period of time there are some differences between PADM2M and Sint-2000 resulting from the selection and interpretation of raw 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 paleomagnetic reconstructions (Morzfeld et al. 2017).

4.2.2 Building on previous work in threshold-based predictions

There have been several studies concerning the behavior of the magnetic field leading up to a reversal or major excursion, in particular, with regards to the decay in intensity of the modern

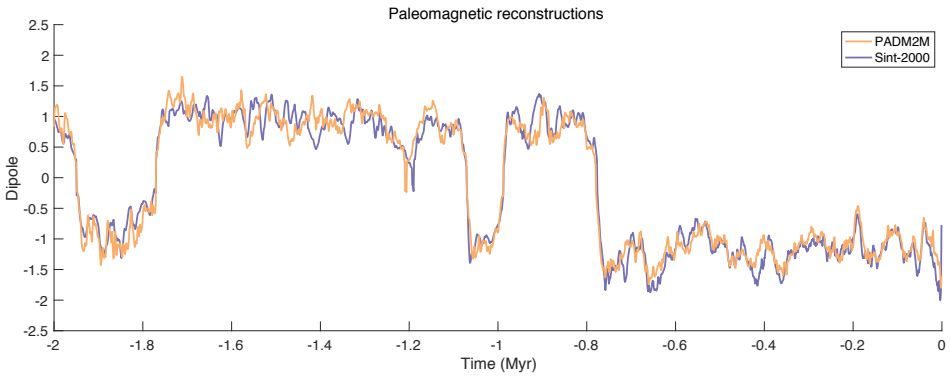

Figure 4.2: The PADM2M (yellow) and Sint-2000 (purple) paleomagnetic reconstructions of the dipole covering the last 2 Myr. The sign indicates polarity and the amplitude is scaled such that the average intensity for each time series is one.

field (see, e.g., Hulot et al. 2002; Constable & Korte 2006; Olson et al. 2009; Laj & Kissel 2015). However, the testing of explicit prediction strategies applied to dipole intensity time series remains limited to the data assimilation study of Morzfeld et al. (2017) and the threshold-based predictions of Gwirtz et al. (2021). We build the search for precursors to reversals on the framework of the threshold-based prediction study and therefore, provide a review of relevant details below.

Low-dipole events

The numerical experiments of section 4.4 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 dipole changes sign or the intensity drops below a valued 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 figure 4.3 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 sub-chron, seen in PADM2M and Sint-2000 around 1.2 Myr in the past (figure 4.2) is a single low-dipole event. To make this definition comparable across the models and reconstructions, dipole time series are scaled to the respective model or

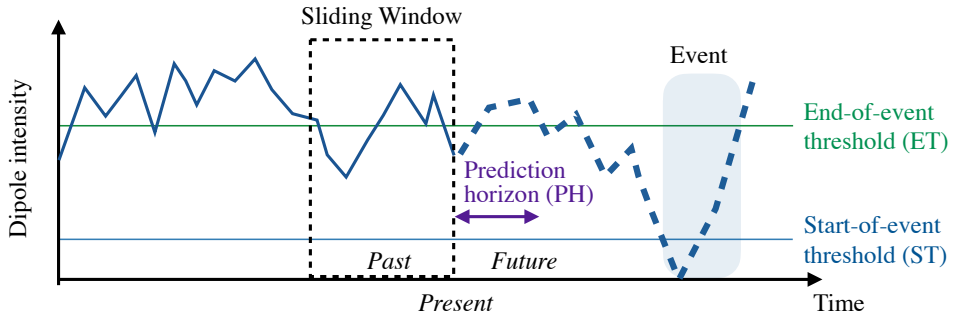

Figure 4.3: Illustration of the prediction strategy. A time series of dipole intensity (thick blue) is shown as a function of time. The thin blue and green horizontal lines show the start-of-event and end-of-event thresholds. The sliding window (dashed rectangle) encompasses the time series segment used to make a prediction of whether a low-dipole event will start within a prediction horizon (purple arrow) from the present (right side of the window).

reconstruction's average intensity and thresholds are expressed as a percentage of that average. For example, figures 4.1 and 4.2 are scaled in this manner and therefore a non-dimensional intensity of one is the long-term average of the model or reconstruction. As we use thresholds of ST= 10% and ET= 80% (following Gwirtz et al. (2021)), events begin when the intensity drops below 0.1 and end when it exceeds 0.8.

Prediction horizons

Predictions are made concerning whether a low-dipole event will begin within an *a priori* specified period of time, called the prediction horizon (PH). This is illustrated in figure 4.3 where the length of the double-sided purple arrow indicates the size of the prediction horizon, i.e., the time interval by which we wish to anticipate the start of a low-dipole event. In the illustration, an event does occur in the future, but beyond the range of the prediction horizon. Therefore, the correct prediction would be that no low-dipole event begins within the prediction horizon.

The prediction horizons need to be chosen prior to any training and validation and should be such that an attempt is being made to anticipate the start of low-dipole events over a meaningful interval of time. Specifically, if a prediction horizon is too short, we may only be predicting

events once they have essentially already begun. Conversely, for an excessively long prediction horizon, anticipating an event becomes trivial because one is likely to occur within a large enough window of time, simply by chance. 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 after it was verified that results where not highly sensitive to choices of other prediction horizons in this range. This allowed for meaningful comparisons in the predictability across a collection of models and reconstructions because the dimensional time of the models (discussed above) are not selected to match the timescales over which reversals and excursions occur. Instead, they largely reflect the timescales of short-term fluctuations. For this reason we choose to use predictions horizons of one average event duration from Gwirtz et al. (2021) shown in table 4.1.

Table 4.1: Prediction horizons used for the models and paleomagnetic reconstructions.

	G12	3-D	PADM2M	Sint-2000
Prediction horizon (PH)	3.2 kyr	16.4 kyr	11.7 kyr	10.2 kyr

Threshold-based predictions

The threshold-based prediction strategy of Gwirtz et al. (2021) is simply, predict a low-dipole event will begin within the prediction horizon whenever the dipole intensity is below a predefined level called the *warning threshold*. At a time a prediction is made, the true outcome is labeled 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 can then be labeled as a true positive (TP), true negative (TN), false positve (FP) or false negative (FN). A warning threshold is determined 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)

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{4.2}$$

A perfect MCC score is one, while a score near zero indicates poor predictions. The MCC is useful as it is robust when evaluating classifications of imbalanced data (Chicco & Jurman 2020) like we face in the prediction of low-dipole events. Specifically, low-dipole events are rare and therefore one could obtain a high *accuracy* by always forecasting no event to occur. The use of the MCC avoids settling on this prediction strategy and for that reason, we use it in this paper.

The threshold prediction strategy has been applied to numerical models, including G12 and the 3-D simulation, as well as the paleomagnetic reconstructions PADM2M and Sint-2000. The skill of the strategy varies widely among this collection with G12 being the most predictable, followed by the paleomagnetic reconstructions and finally the 3-D simulation (Gwirtz et al. 2021). We apply the threshold strategy, alongside machine learning predictions, to determine a baseline for performance.

4.2.3 Machine learning methods

We primarily use linear SVMs to search for precursors to reversals and major excursions. This is done by training them to classify segments of time series according to whether they precede low-dipole events (see section 4.3). We use SVMs because they have been around for nearly three decades (Cortes & Vapnik 1995) and have since become a fundamental and widely used machine learning technique (see, e.g, Hamel 2011; Ma & Guo 2014; Murty & Raghava 2016). Additionally, linear SVMs are conceptually, easy to understand. In simple terms, training a linear support vector machine 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).

We acknowledge the potential of results to be sensitive to the particular machine learning approach chosen. In particular, the failure of a machine learning model to find precursors to reversals cannot be taken as conclusive evidence of the absence of such precursors. For this reason we conclude the study in section 4.5.3, with a limited number of experiments testing the robustness of particular results using *long short-term memory* networks (LSTMs, Hochreiter & Schmidhuber 1997) and non-linear SVMs. LSTMs are a form of recurrent neural network which are often used for classifying time series (see, e.g., Graves 2012). Non-linear SVMs effectively transform data, which may not be linearly separable, to a higher-dimensional space where it is potentially, linearly separable (see, e.g., Cristianini et al. 2000).

All machine learning methods we employ use functions from Matlab's machine learning toolbox. Due to the imbalanced nature of the data mentioned in section 4.1, we make a small adjustment to the standard algorithms and the procedures typically used to train models. The details of these modifications are discussed below in section 4.3.

4.3 Training SVMs for imbalanced data

The setup for making predictions using machine learning is as follows. At a given point in a time series of dipole intensity, we examine the recent past to search for precursors to low-dipole events and make a prediction. Predictions are only made concerning whether a low-dipole event will begin within the prediction horizon (see section 4.2.2). This is illustrated in figure 4.3 where the sliding window (dashed rectangle) outlines 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 prediction horizon. For numerical experiments, we do not consider windows during which an event is occurring. Simply put, we do not make predictions once an event has started or, if the window of past history overlaps with the end of a prior event. As with the threshold-based predictions described in section 4.2.2, for each prediction the "true" outcome is labeled either, Postive (P), or Negative (N), depending on whether a low-dipole event begins within the prediction horizon. In this way, the challenge of making predictions is a binary classification problem (Goodfellow et al. 2016). To put it simply, we examine the segment of a dipole time series within the sliding window and attempt to determine whether it belongs to the class P or N, i.e., whether an event will soon begin or not. For example, at the time a prediction is being made in figure 4.3, the correct classification for the time series segment in the sliding window is N; an event starts in the future but not within the prediction horizon. Then, just as with threshold strategy, every prediction (classification) results in either a TP, TN, FP or FN. Given this, we we follow Gwirtz et al. (2021) and use the MCC to evaluate predictions.

Machine learning models however, are generally not designed to maximize MCC but to minimize a particular loss function. During training, the SVM model we use is designed to minimize a loss function which measures the extent of separation of classes, and the extent of misclassifications, by a hyperplane (see section 4.2.3). Because of this, we determine "optimal" SVMs for each particular size of sliding window (how much recent history is considered in making a prediction) in the following way. In each numerical experiment we perform, a time series or a set of time series is divided in various ways into separate *training* and *validation* data. The training and validation data are independent of one another and only the training data are used in determining the optimal SVM which is used to make predictions with the validation data. We provide the training data to an SVM and reduce by varying degrees, the weight given to misclassified negatives (false positives) in the loss function. In effect, this emphasizes the need to correctly classify the rare positives (P) over the more common negatives and thus, increases

MCC. We then select as optimal, the SVM which maximizes MCC on the training data. Figure 4.4 shows both accuracy (open circles) and the MCC (closed circles) achieved during the training of an SVM on a short G12 simulation, as a function of the relative size of penalty applied to false positives compared to false negatives (FP/FN penalty ratio). Notice that all SVMs achieve

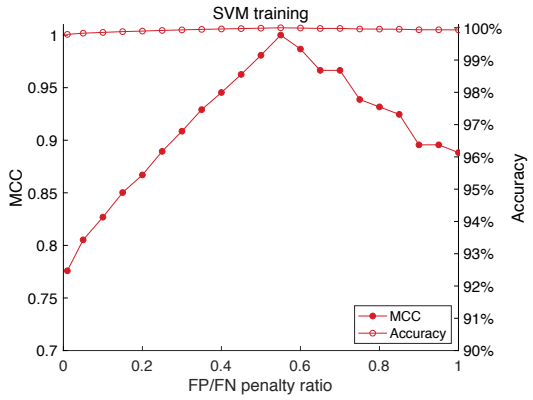

Figure 4.4: Accuracy (open circles) and MCC (closed circles) as a function of the relative size of penalty applied to false positives compared to false negatives, during SVM training.

accuracy near 100% during training—as would any strategy that only rarely predicts low-dipole events to occur. The MCC scores however, vary between 0.77 and 1 and we select the SVM trained with the FP/FN penalty ratio of 0.55 as optimal because it achieves the highest MCC.

4.4 SVM prediction results

Results from numerical experiments with SVMs applied to the models and paleomagnetic reconstructions are presented. All results use the prediction horizons of table 4.1. In each experiment we also report the resulting skill from applying the threshold strategy (see section 4.2.2), using the same training and validation data as the SVM. We begin with results from the numerical models followed by the paleomagnetic reconstructions.

4.4.1 Results with the G12 and 3-D numerical models

SVMs trained on long time series

We begin by examining the numerical models under the ideal circumstances where one has a long training and validation time series each containing a large number of events. In this case, each of the two G12 time series contains 180 events. For the 3-D simulation, the axial dipole time series is divided into separate training and validation data sets containing 180 events each. For both models and each window size considered, the SVM which maximizes MCC on the training data is applied to the validation data. The scores on the validation data as a function of the window size used are shown by the red (G12) and green (3-D) circles of figure 4.5. The

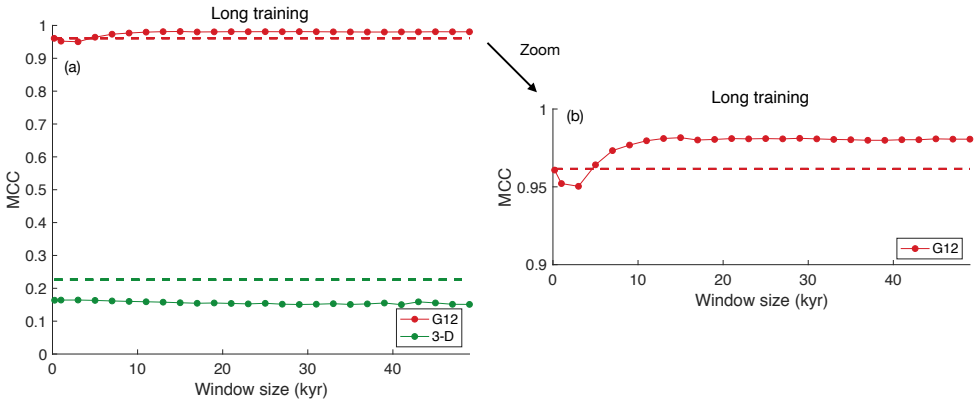

Figure 4.5: Validation MCC as a function of sliding window size for SVMs trained on a large data set from G12 (red circles) and the 3-D model (green circles). Horizontal dashed lines indicate the MCC achieved by the threshold strategy using the same training and validation data from G12 (red) and 3-D (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).

dashed horizontal lines indicate the MCC achieved by a simple threshold strategy with the same sets of training and validation data. With both the threshold strategy and the SVMs, we find that events in G12 are fairly predictable (MCC near one) while the prediction of events in the 3-D model is more challenging (MCC closer to zero).

We note that for G12, the SVM performs consistently better than the threshold strategy with window sizes greater than 10 kyr (see the magnified plot of panel (b) in figure 4.5). This

indicates that the SVMs are identifying features in the recent history of G12 time series which signal oncoming low-dipole events. The fact that the MCC score stabilizes beyond window sizes of around 11 kyr for the G12 model, suggests that the SVMs are not finding additional information signaling a low-dipole event when looking beyond 11 kyr into the past. This implies that a window size of around 11 kyr is optimal for G12 in the sense that, the SVMs which look further back in time gain no advantage. Note that the stabilizing of the MCC for large windows is to be expected—the state of the dipole, if we go far enough into the past, should be unrelated to its present and future behavior.

Unlike the G12 model, predictions of the 3-D simulation do not improve as the window size is increased. Indeed the curve of MCC scores for the 3-D simulation is flat—the validation skill score remains essentially unchanged regardless of window size. This indicates that unlike with the G12 model, the SVMs do not benefit from accessing the recent history of the axial dipole of the 3-D simulation. We remind the reader that this does not mean that precursors to low-dipole events do not exist in the 3-D time series, only that the SVM has failed to find any.

SVMs trained on short time series

We note that while the data used to train the SVMs of figure 4.5 is large, the observational record of the Earth's field which can be used for studies such as this is relatively short. The paleomagnetic reconstructions of PADM2M and Sint-2000 which we examine in section 4.4.2 span just the last 2 Myr and resolve 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 remains the same). The resulting validation MCC is shown as a function of window size by the red (G12) and green (3-D) circles of figure 4.6. The dashed lines indicate the validation MCC for the threshold strategy trained on the same, short time series containing only five events. In Gwirtz et al. (2021) it was shown that a useful threshold could be determined for G12 from a training data set containing only five events. This is evident in figure 4.6 by the

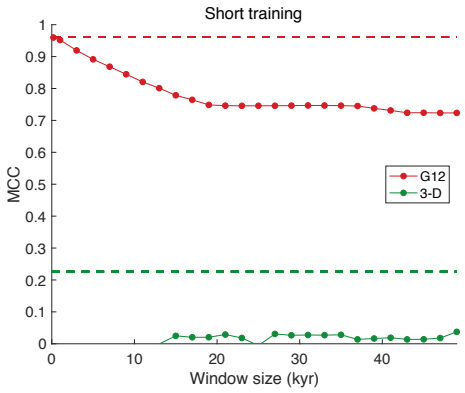

Figure 4.6: Validation MCC as a function of sliding window size for SVMs trained on a small data set.

the dashed lines as well as the SVM validation score for a window size of one time step in G12. However, for the 3-D simulation, all SVMs trained on the short time series for window sizes 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 larger window sizes the validation skill scores of SVMs trained on G12 steadily drop off. Larger window sizes should only have the potential to add information relative to shorter ones and therefore the drop in skill score with window size suggests the SVMs are overfitting to the short training data. Perhaps unsurprisingly, in the less predictable 3-D time series we observe similar behavior. With large windows, validation MCC scores for the 3-D simulation are notably lower than when training with large data. Indeed the MCC scores even drop below zero, indicating one is at times better predicting the opposite of the trained SVM.

4.4.2 SVMs with paleomagnetic reconstructions

Studying SVMs with the PADM2M and Sint-2000 paleomagnetic reconstructions requires a change in approach due to the shortness of the two time series. Specifically, we modify the method of selecting training and validation data. Instead of dividing each time series 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 window size and prediction

horizon, all of the segments of the time series to be used for either training or validation are collected and classified as positives or negatives (see section 4.3). 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. Figure 4.7 summarizes the results of this process. For each window size the closed

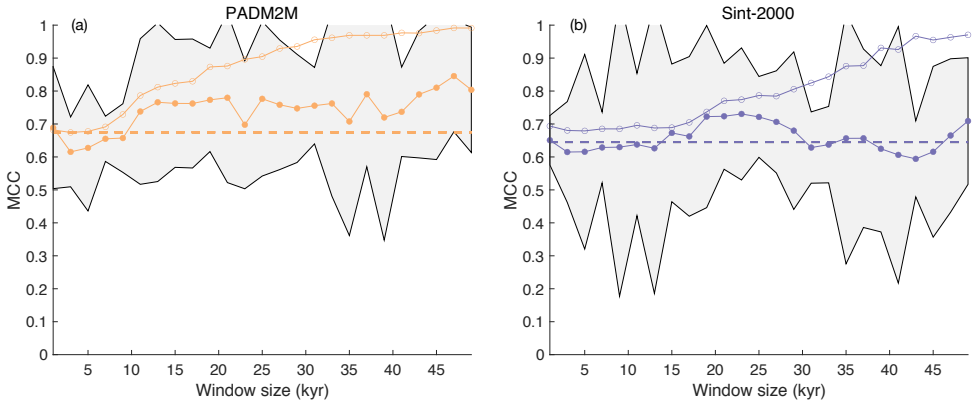

Figure 4.7: Average MCC (closed circles) with two standard deviations (grey cloud) as a function of sliding window size resulting from stratified 5-fold cross validation with (a) PADM2M and (b) Sint-2000. The dashed lines show the average MCC of the threshold strategy subject to the same stratified 5-fold cross validation. Open circles show the training MCC achieved by fitting to the full data.

circles record the average of the five MCC validation scores with the grey region region indicating the average plus/minus two standard deviations. The open circles show the training MCC from fitting an SVM to the full time series and the dashed lines report the average MCC score resulting from following the same training and validation procedure with the threshold strategy.

With the results of figure 4.7 we ask: are the SVMs detecting a signal in the observational record of the dipole which indicates oncoming low-dipole events? Unfortunately the limited data leads to large variations in the SVM validation skill scores and therefore any interpretation of the results must be done under this uncertainty. Additionally, as discussed in section 4.2.1, there is uncertainty in the paleomagnetic reconstructions themselves, as evidenced by the differences

in PADM2M and Sint-2000. These differences 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 figure 4.5) which would suggest the existence of a signal preceding and event. However, this is not the case with Sint-2000 where on average, the validation skill score seems to remain unchanged regardless of window size (similar to the 3-D results of figure 4.5).

4.5 Discussion

In this section we discuss the results of the numerical experiments for both the numerical simulations and paleomagnetic reconstructions. We then investigate the sensitivity of the results to the choice of machine learning technique.

4.5.1 Discussion of experiments with SVMs and numerical models

The skill of SVMs in making predictions of the numerical models differs significantly between G12 and 3-D simulations. When trained on large data, SVM predictions of G12 are of high quality (MCC near 1) while predictions of the 3-D simulation are poor (MCC of around 0.2). This difference in predictability is in line with results using a threshold-based strategy (Gwirtz et al. 2021). When trained on small data sets containing only five low-dipole events, the skill of SVMs deteriorate with an increased window size. This suggests that when using limited data, such as is available in paleomagnetic reconstructions, linear SVMs tend to overfit during training and it is better to use a simple threshold strategy.

What is perhaps most notable however, is not the particular MCC values but the the shape of the SVM curves in figure 4.5. When training on long simulations, predictions of G12 improve with an increased window size, indicating that dipole intensity time series of G12 contain precursors which signal oncoming low-dipole events. This is in agreement with Morzfeld et al.

(2017) where it was found that predictions with the G12 model were improved when assimilating data over a window of time. Conversely, the curve of validation MCC scores for predictions of the 3-D simulation in figure 4.5 is flat, showing no such improvement with window size and indicating that the linear SVMs do not find precursors to low-dipole events. We remind the reader that the G12 model is defined by a system of ordinary differential equations for three scalers, one of which represents the axial dipole (see section 4.2.1). Therefore, when SVMs identify precursors to low-dipole events in a dipole intensity time series of G12, they are finding indicators that the unobserved components of the system (the two other scaler values) are in a state favorable for causing a low-dipole event.

If we consider low-dimensional models which are defined only by a single scaler, precursors to low-dipole events may not exist by construction. For example consider stochastic differential equation (SDE) models of the form

$$\mathrm{d}x = f(x)\mathrm{d}t + \sqrt{2q}\,\mathrm{d}W, \tag{4.3}$$

where $f(x)$ is a prescribed function of x, q is a constant, W is Brownian motion and the dipole intensity is a deterministic function of x. Notice that if the increments of Brownian motion are uncorrelated in time, the only information useful to predicting future behavior is the present value of $f(x)$. The SDE models of Morzfeld & Buffett (2019) and Pétrélis et al. (2009) take the form of equation (4.3). We consider them under the parameter regimes used in Gwirtz et al. (2021) and refer to them as DW (Morzfeld & Buffett 2019) and P09 (Pétrélis et al. 2009). In figure 4.8 we show the result of performing linear SVM experiments using long training and validation simulations (180 events each) with the DW model (dark purple) and P09 model (blue). As in section 4.4, the circles represent validation MCC scores from linear SVMs while the dashed lines indicate the validation score of the threshold strategy. For both models, the SVM scores do not improve with a larger window, just as was seen with the 3-D simulation. This is to be expected

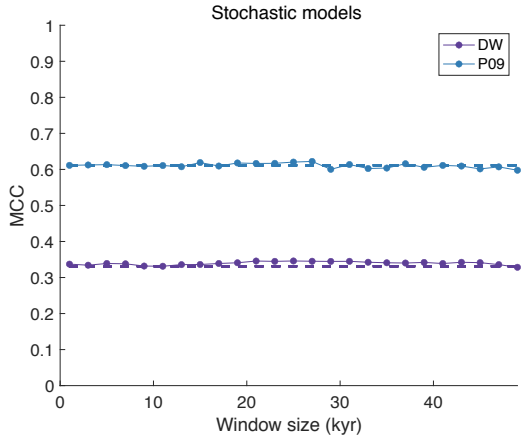

Figure 4.8: Validation MCC as a function of sliding window size for SVMs trained on a large data set from DW (dark purple circles) and P09 (blue circles). Horizontal dashed lines indicate the MCC achieved by the threshold strategy using the same training and validation data from DW (dark purple) and P09 (blue).

because both DW and P09 are constructed such that, except possibly *during* the occurrence of, and recovery from a low-dipole event, the dipole intensity indicates the value of $f(x)$ and therefore, no additional information is obtained by looking into the past. That the 3-D simulation exhibits the same behavior supports the use of SDE models of this type as appropriate low-dimensional models.

4.5.2 Discussion of experiments with SVMs and paleomagnetic reconstructions

The limited data of the paleomagnetic reconstructions of PADM2M and Sint-2000 makes them difficult to interpret with regards to the machine learning studies of this paper. The linear SVM experiments of section 4.4.2 do not show convincing evidence of precursors to reversals in the paleomagnetic record. Additionally, it must be recognized that any signal which might be found in time series of the paleomagnetic reconstructions would not necessarily reflect a property of the Earth's 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, and the mathematical methods

used to fit that record. If there are indeed no precursors to low-dipole events in the paleomagnetic reconstructions, this supports the use of SDE models such as DW and P09 with the same property (see above, section 4.5.1). 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, the results suggest the the threshold strategy may be the best one can do when making predictions based solely on Earth's dipole intensity however, the value of that strategy has been shown to be limited.

4.5.3 Sensitivity of results to the choice of machine learning model

We investigate the sensitivity of the results to the choice of linear SVMs. In particular, we test whether the use of linear SVMs has led us to miss precursors to low-dipole events. First, we directly analyze the time series of the various simulations and paleomagnetic reconstructions. 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 dipole intensity. In simple terms, if the ith entry of a dipole intensity time series is d_i, we compute a time series with entries $\Delta d_i = d_{i+1} - d_i$. The autocorrelation function of the time series of change in dipole intensity is shown for both models and paleomagnetic reconstructions in figure 4.9. The intention is to determine if, at any given time, correlations exist between the change that is about to occur in dipole intensity, and previous changes. For the purpose of making comparisons we use resampled time series of simulations so that the temporal resolution is approximately equal to the 1 kyr resolution of the paleomagnetic reconstructions. We see that G12 exhibits the strongest such correlations, with the the change in 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 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 PADM2M, which also indicate, along with Sint-2000, some correlation between

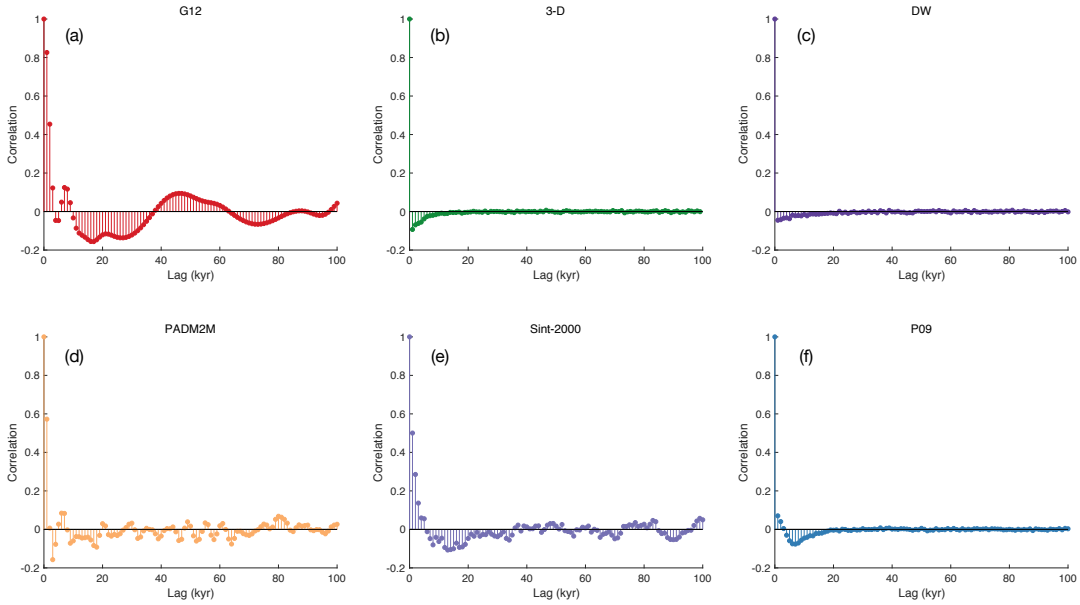

Figure 4.9: Autocorrelation functions of time series of change in dipole intensity for (a) G12, (b) 3-D, (c) DW, (d)PADM2M, (e) Sint-2000 and (f) P09.

future changes and recent history. Interestingly, but perhaps unsurprisingly given the results of the SVM experiments, changes in the dipole intensity of the 3-D model are largely uncorrelated over timescales of 1 kyr. Indeed, the autocorrelations of the 3-D simulation appear most similar to the SDE models of DW and P09 for which we know future behavior is independent of the past. This supports the conclusion that the dipole of the 3-D simulation does not contain precursors to low-dipole events.

We further test the results by repeating a limited set of experiments using a non-linear SVM. Specifically, we use an SVM with radial basis functions (RBFs, see, e.g., Cristianini et al. 2000). The top row of figure 4.10 shows the result of applying this SVM model to the paleomagnetic reconstructions using the same stratified, k-fold cross validation of section 4.4.2. The closed circles show the average validation MCC of the SVMs with the grey cloud indicating two standard deviations. The open circles show the training MCC resulting from fitting a non-linear SVM to the full data set. Dashed lines show the average validation MCC of the threshold strategy subject to the same stratified 5-fold cross validation. Although the increasing skill with

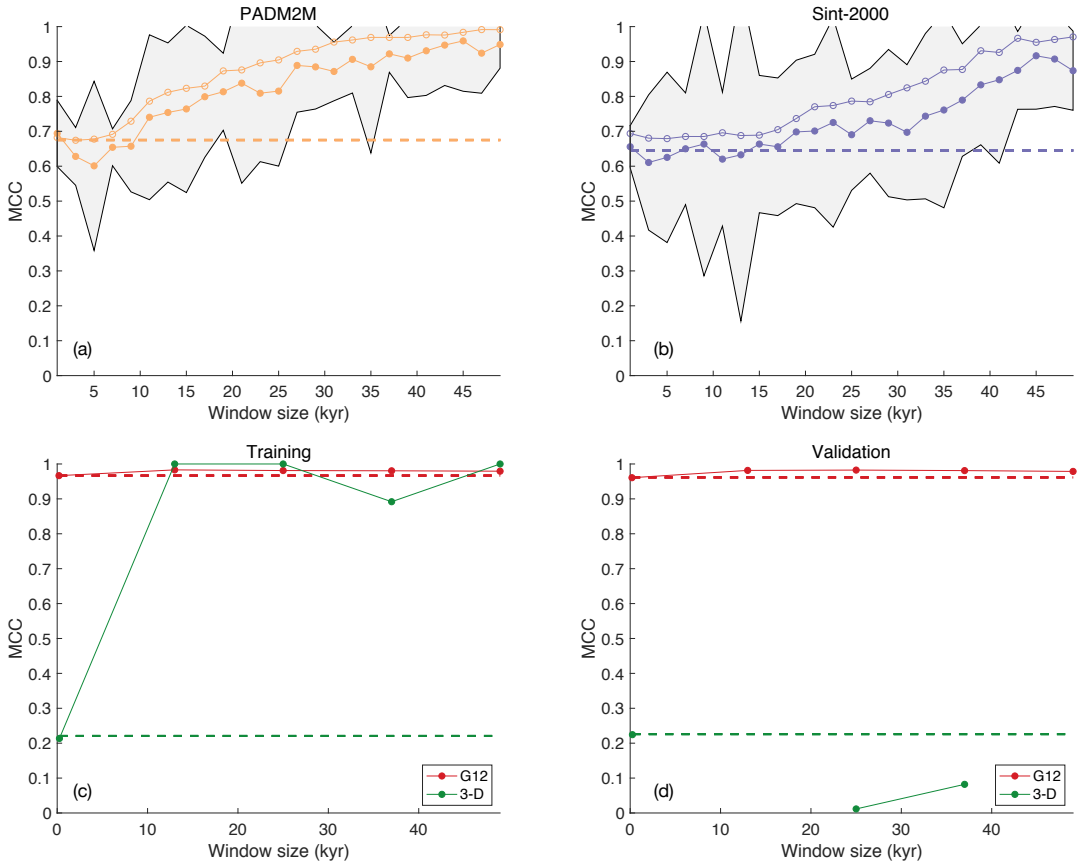

Figure 4.10: Top row: Average MCC (closed circles) with two standard deviations (grey cloud), as a function of sliding window size, resulting from stratified 5-fold cross validation with (a) PADM2M and (b) Sint-2000 using SVMs with RBFs. The dashed lines show the average MCC of the threshold strategy subject to the same stratified 5-fold cross validation. Open circles show the training MCC achieved by fitting to the full data. Bottom row: (c) Training and (d) validation scores using long simulations of G12 (red) and 3-D (green) with non-linear SVMs (circles) and the threshold strategy (dashed lines).

increasing window size suggests the possibility that the non-linear SVMs are finding precursors to low-dipole events, it becomes evident upon further examination that the RBFs allow for a level of overfitting which makes these results unreliable. Specifically, the bottom row of figure 4.10 shows training and validation scores using non-linear SVMs (RBFs) with long simulations of the G12 and 3-D model. The results of G12 are similar to those of figure 4.5 but we see that with the 3-D simulation, the SVM obtains training MCCs near 1. Applying those SVMs to the validation data, it becomes clear that the model has overfit, with several failing to predict any events (MCC

is undefined and thus, not reported).

Next 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), i.e., stratified k-fold cross validation with the paleomagnetic reconstructions and large training and validation data sets with the numerical models. The results are shown in figure 4.11 where we again find no clear evidence of low-dipole event precursors in the paleomagnetic reconstructions (top row). Indeed, with each reconstruction, for multiple window sizes there are

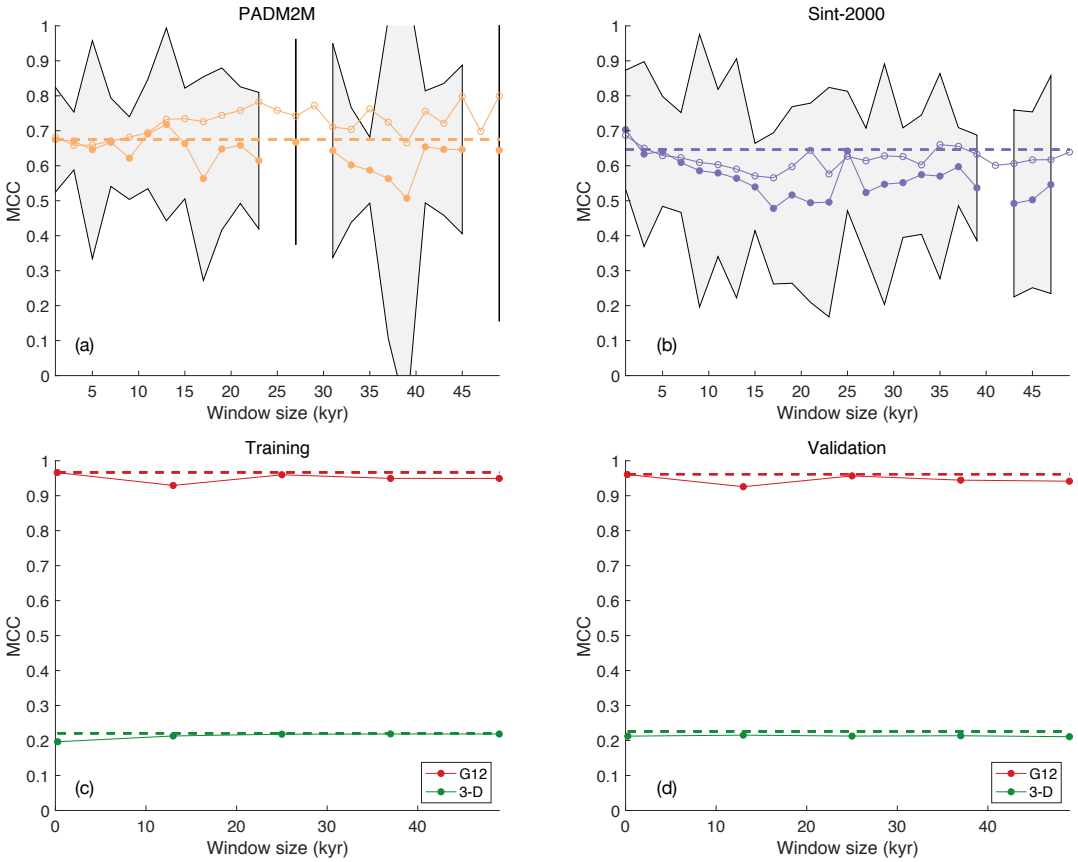

Figure 4.11: Top row: Average MCC (closed circles) with two standard deviations (grey cloud) as a function of sliding window size resulting from stratified 5-fold cross validation with (a) PADM2M and (b) Sint-2000 using LSTMs. Windows sizes for which there are no values had validation runs where no events were predicted, resulting in an undefined MCC. The dashed lines show the average MCC of the threshold strategy subject to the same stratified 5-fold cross validation. Open circles show the training MCC achieved by fitting to the full data. Bottom row: (c) Training and (d) validation scores using long simulations of G12 (red) and 3-D (green) with LSTMs (circles) and the threshold strategy (dashed lines).

validation runs which predict no events, resulting in undefined MCC scores and thus no reported averages or standard deviations (no closed circles or grey cloud). For the numerical models (bottom row) G12 consistently results in MCC scores near one however, we fail to find the same consistent improvement with window size seen with SVMs. This is perhaps due to the difficulty of tuning the FP/FN penalty ratio of the LSTM carefully enough to obtain a marginal improvement in an already high MCC score. Indeed if we increase the difficulty of predicting low-dipole events in G12 by selecting a larger prediction horizon, we find that the LSTM obtains a qualitatively similar results to the linear SVM. Specifically, the predictions consistently improve with increased window size indicating that the LSTMs, like the SVMs, find precursors to low-dipole events in G12. Most importantly, we see that results with the 3-D simulation are unaffected by window size. This further supports the conclusion that the dipole intensity time series of the 3-D simulation contains no precursors to low-dipole events.

4.6 Concluding remarks

The objective of this study was to search for precursors to reversals and major excursions in time series of dipole intensity and investigate their value in predicting reversals and major excursions. Previous efforts to predict such events used data assimilation (Morzfeld et al. 2017) and a threshold strategy (Gwirtz et al. 2021). In this study, we used machine learning models to examine segments of dipole intensity time series and make predictions. Given that the observational record is limited while machine learning models generally rely on large data sets for independent training and validation, we studied simulations of numerical models in addition to paleomagnetic reconstructions. The framework of the threshold study of Gwirtz et al. (2021) was largely adopted to define low-dipole events and evaluate predictions. In particular, because low-dipole events occur infrequently, we used the MCC to evaluate machine learning models because it is a more robust measure of success in predicting rare events than, for example, accuracy (see

figure 4.4). Studying machine learning models using MCC as a measure of success required a small modification to the standard training processes (see section 4.3). We focused on the use of linear SVMs for machine learning models but also tested the robustness of results with non-linear SVMs and LSTMs.

We found that the skill of machine learning in making predictions varied among the dipole models and reconstructions. Only in the dipole of the G12 model did we find clear evidence of precursors to reversals and major excursions. Experiments with the dipole of a reversing 3-D geodynamo simulation indicate that it possesses no such precursors to low-dipole events. With both numerical models, it was found that prediction performance deteriorated when training on short simulations containing only five low-dipole events, suggesting that it may be difficult to train a useful machine learning model on the limited paleomagnetic record. When experiments were performed with paleomagnetic reconstructions (PADM2M and Sint-200), no convincing evidence of low-dipole event precursors was found. The findings with the 3-D simulation and paleomagnetic reconstructions support the use of low-dimensional SDE dipole models for which future behavior has little to no dependence on the past history.

4.6.1 Outlook

It cannot be ruled out that other techniques could find precursors to low-dipole events in the paleomagnetic reconstructions and dipole intensity time series of the 3-D simulation. More sophisticated machine learning models however, require a greater amount of training data to constrain a larger set of model parameters. It is therefore unlikely that machine learning techniques of increased complexity would identify with confidence, precursors in PADM2M or Sint-2000. Consideration should be given to supplying machine learning models with other features of the 3-D simulation, in addition to the axial dipole. This could help to better identify and understand the dynamics which lead to reversals and major excursions. Relating results to the Earth however, would require careful consideration in view of the limitations to our knowledge

of the past and even present state of the geodynamo, outside of the large-scale features of the magnetic field.

Acknowledgements

KG acknowledges that this work was supported by NASA Headquarters under the NASA Earth and Space Science Fellowship Program - Grant "80NSSC18K1351".

Chapter 4, is currently being prepared for submission for publication of the material. Gwirtz, K., Davis, T. and Morzfeld, M., Can machine learning find precursors to reversals of Earth's magnetic dipole field? The dissertation author was the primary investigator and author of this material.

References

Buffett, B., 2015. Dipole fluctuations and the duration of geomagnetic polarity transitions, *Geophysical Research Letters*, **42**, 7444–7451.

Buffett, B. & Matsui, H., 2015. A power spectrum for the geomagnetic dipole moment, *Earth and Planetary Science Letters*, **411**, 20–26.

Buffett, B., Ziegler, L., & Constable, C., 2013. A stochastic model for paleomagnetic field variations, *Geophysical Journal International*, **195**(1), 86–97.

Buffett, B. A., King, E. M., & Matsui, H., 2014. A physical interpretation of stochastic models for fluctuations in the earth's dipole field, *Geophysical Journal International*, **198**(1), 597–608.

Cande, S. & Kent, D., 1995. Revised calibration of the geomagnetic polarity timescale for the late cretaceous and cenozoic, *Journal of Geophysical Research: Solid Earth*, **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 Genomics*, **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, *Machine learning*, **20**(3), 273–297.

Cristianini, N., Shawe-Taylor, J., Shawe-Taylor, D., Books24x7, I., & Press, C. U., 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press.

Gissinger, C., 2012. A new deterministic model for chaotic reversals, *Eur. Phys. J. B.*, **85**, 137.

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?, *Geophysical Journal International*, **225**(1), 277–297.

Hamel, L., 2011. *Knowledge Discovery with Support Vector Machines*, Wiley Series on Methods and Applications in Data Mining, Wiley.

Hochreiter, S. & Schmidhuber, J., 1997. Long Short-Term Memory, *Neural Computation*, **9**(8), 1735–1780.

Hoyng, P., Ossendrijver, M., & Schmitt, D., 2001. The geodynamo as a bistable oscillator, *Geophysical and Astrophysical Fluid Dynamics*, **94**, 263–314.

Hulot, G., Eymin, C., Langlais, B., Mandea, M., & Olsen, N., 2002. Small-scale structure of the Geodynamo inferred from Oersted and Magsat satellite data, *Nature*, **416**, 620–623.

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

Laj, C. & Kissel, C., 2015. An impending geomagnetic transition? Hints from the past, *Front. Earth Sci.*, **3**, 61.

Lhuillier, F., Hulot, G., & Gallet, Y., 2013. Statistical properties of reversals and chrons in numerical dynamos and implications for the geodynamo, *Physics of the Earth and Planetary Interiors*, **220**, 19–36.

Lowrie, W. & Kent, D., 2004. Geomagnetic polarity time scale and reversal frequency regimes, *Timescales of the paleomagnetic field*, **145**, 117–129.

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

Meduri, D. & Wicht, J., 2016. A simple stochastic model for dipole moment fluctuations in numerical dynamo simulations, *Frontiers in Earth Science*, **4**, 38.

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

Morzfeld, M., Fournier, A., & Hulot, G., 2017. Coarse predictions of dipole reversals by low-dimensional modeling and data assimilation, *Physics of the Earth and Planetary Interiors*, **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 time scale 2012*, chap. 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, *Physics of the Earth and Planetary Interiors*, **173**(1), 121–140.

Olson, P., Deguen, R., Hinnov, L. A., & Zhong, S., 2013. Controls on geomagnetic reversals and core evolution by mantle convection in the phanerozoic, *Physics of the Earth and Planetary Interiors*, **214**, 87–103.

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

Schmitt, D., Ossendrijver, M., & Hoyng, P., 2001. Magnetic field reversals and secular variation in a bistable geodynamo model, *Phys. Earth planet. Inter.*, **125**, 119–124.

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, *Geophysical Journal International*, **184**(3), 1069–1089.

Chapter 5

Concluding comments

5.1 Summary

The purpose of this dissertation was to develop tools for understanding and predicting the dynamic behavior of the geodynamo over decadal and millennial timescales. There are both practical and scientific motivations for this pursuit. In practical terms, forecasts of the Earth's magnetic field play an important role in, e.g, navigation, surveying and the planning of satellite deployments. From a purely scientific perspective, variations in the Earth's magnetic field provide one of the few windows we have into the dynamics of the planet's deep interior. We took a variety of approaches to the study of this topic.

In chapter 2, we focused on the growing field of geomagnetic data assimilation (DA), where information on the magnetic field is assimilated into a numerical geodynamo for the purpose of estimating the dynamic state of the core and producing decadal-scale forecasts of the magnetic field. While geomagnetic DA has shown promise, there remain important questions about how it can be most effectively implemented. Following the path of DA development in other applications, we constructed a proxy model for exploring some of those questions. The proxy model was then used in a large collection of numerical experiments which emphasized the need for localization and inflation in geomagnetic DA, and suggested improvements to current geomagnetic DA schemes.

In chapters 3 and 4, we investigated millennial timescale variations in the Earth's magnetic field. Specifically, we focused on characterizing and predicting reversals and major excursions of Earth's magnetic dipole using time series of dipole intensity. We started by developing an explicit definition of the *low-dipole events* (reversals and major excursions) that we wished to study. Given the limited availability of detail on low-dipole events in the observational record, we used a collection of numerical models in addition to paleomagnetic reconstructions. First we employed a simple threshold strategy to make predictions of low-dipole events (chapter 3). It was found that the predictability varied widely among the models and reconstructions with the results highlighting differences in the way low-dipole events occur. In particular, in the paleomagnetic reconstructions and select low-dimensional models, dipole intensity tends to recover from low values faster than it decays to them. The opposite however, was found in the dipole of a low-dimensional model and a reversing 3-D numerical geodynamo. Next we searched for precursors to low-dipole events in dipole intensity time series, by using machine learning models (chapter 4). To do this, we trained support-vector machines and long short-term memory networks to examine segments of dipole intensity time series and determine whether they precede an oncoming low-dipole event. We found no convincing evidence of precursors to low-dipole events in the dipole intensity of paleomagnetic reconstructions or a reversing 3-D numerical geodynamo. The results of the study of reversals and major excursions led to the suggestion of new criteria for evaluating how Earth-like the dipole of a numerical model is.

5.2 Outlook

The immediate plan for building on the work of this dissertation concerns geomagnetic DA and the contents of chapter 2. Specifically, the dissertation author will use the proxy model of chapter 2, in conjunction with NASA's Geomagnetic Ensemble Modeling System (GEMS, Tangborn et al. 2021), to further develop and validate new tools for the next generation of

geomagnetic DA systems. The EnKF-based GEMS is NASA's operational geomagnetic DA system. Similar to the study of chapter 2, work will concern localization and inflation strategies, but with the added complication of systematic model bias.

Largely due to computational constraints, current numerical geodynamo models operate in parameter regimes that differ significantly from the Earth (Christensen et al. 2010). For example, many simulations use an Ekman number E (ratio of viscous and magnetic diffusion timescales) in the range of $E \sim 10^{-7}$ (see, e.g., Aubert et al. 2017) while the value for Earth's core is believed to be $E \sim 10^{-15}$ (Braginsky & Roberts 1995). Such differences in parameters almost certainly lead to large, systematic model biases in numerical dynamo simulations which likely have an adverse impact on the performance of geomagnetic DA systems. Localization and inflation, in the presence of systematic model bias, will be studied through observing system simulation experiments (OSSEs) with both the proxy model and GEMS. For this purpose, the parameters of the nature run of the OSSEs will differ from those of the model which is used in assimilations (see section 2.3.4 for a discussion of OSSEs). The optimal methodology determined by the OSSEs will then be used to assimilate geomagnetic field models and produce new estimates of the dynamic state of the outer core.

References

Aubert, J., Gastine, T., & Fournier, A., 2017. Spherical convective dynamos in the rapidly rotating asymptotic regime, *J. Fluid Mech.*, **813**, 558–593.

Braginsky, S. I. & Roberts, P. H., 1995. Equations governing convection in Earth's core and the geodynamo, *Geophysical & Astrophysical Fluid Dynamics*, **79**(1-4), 1–97.

Christensen, U. R., Aubert, J., & Hulot, G., 2010. Conditions for Earth-like geodynamo models, *Earth and Planetary Science Letters*, **296**(3), 487–496.

Tangborn, A., Kuang, W., Sabaka, T. J., & Yi, C., 2021. Geomagnetic secular variation forecast using the NASA GEMS ensemble Kalman filter: A candidate SV model for IGRF-13, *Earth, Planets Space*, **73**(47).