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

Gaussian Processes Autoregressive Models for Forecasting the Disturbance Storm Time Index

M. Chandorkar[1], E. Camporeale[1], and S. Wing[2]

[1] Centrum Wiskunde Informatica (CWI), Amsterdam, 1098XG Amsterdam
e-mail: m.h.chandorkar@cwi.nl e.camporeale@cwi.nl
[2] The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland, 20723, USA

ABSTRACT

We present two models, based on Gaussian Processes, for the *One Step Ahead* (OSA) prediction of the *Dst* geomagnetic activity index. The models are Auto regressive, with or without eXogenous inputs (GP-ARX and GP-AR, respectively). We compare the performance of these models with the current state of the art in one step ahead *Dst* prediction on a set of 63 benchmark storms from 1998-2006, previously analyzed in an earlier study. We show that, despite its lack of sophistication, the so-called Persistence model represents an important test to evaluate the performance of a one hour ahead *Dst* model. Contrary to the state-of-the-art models compared in the literature, our models consistently outperform the Persistence model and represent a substantial improvement in the field, when evaluated on standard metrics. Finally, an important feature of the new models is that they naturally provide confidence intervals for their forecast.

Key words. Geomagnetic indices – *Dst* OSA Prediction – Gaussian Processes – Machine Learning

1. Introduction

The magnetosphere's dynamics and its associated solar wind driver form a complex dynamical system. It is therefore instructive and greatly simplifying to use representative indices to quantify the state of geomagnetic activity.

Geomagnetic indices come in various forms, they may take continuous or discrete values and may be defined with varying time resolutions. Their values are often calculated by averaging or combining a number of readings taken by instruments around the Earth. Each geomagnetic index is a proxy for a particular kind of phenomenon. Some popular indices are the K_p, *Dst* and the *AE* index.

1. K_p: The Kp-index is a discrete valued global geomagnetic storm index and is based on 3 hour measurements of the K-indices (Bartels and Veldkamp, 1949). The K-index itself is a three hour long quasi-logarithmic local index of the geomagnetic activity, relative to a calm day curve for the given location.

2. *AE*: The Auroral Electrojet Index, *AE*, is designed to provide a global, quantitative measure of auroral zone magnetic activity produced by enhanced Ionospheric currents flowing below and within the auroral oval (Davis and Sugiura, 1966). It is a continuous index which is calculated every hour.

3. *Dst*: A continuous hourly index which measures the weakening of the Earths magnetic field due to ring currents and the strength of geomagnetic storms (Dessler and Parker, 1959).

For the present study, we focus on prediction of the hourly *Dst* index which is a straightforward indicator of geomagnetic storms. More specifically, we focus on the *one step ahead* (OSA), in this case one hour ahead prediction of *Dst* because it is the simplest model towards building long term predictions of geomagnetic response of the Earth to changing space weather conditions.

The *Dst* OSA prediction problem has been the subject of several modeling efforts in the literature. One of the earliest models has been presented by Burton et al. (1975) who calculated $Dst(t)$ as the solution of an *Ordinary Differential Equation* (ODE) which expressed the rate of change of $Dst(t)$ as a combination of two terms: decay and injection $\frac{dDst(t)}{dt} = Q(t) - \frac{Dst(t)}{\tau}$, where Q(t) relates to the particle injection from the plasma sheet into the inner magnetosphere.

The Burton et al. (1975) model has proven to be very influential particularly due to its simplicity. Many subsequent works have modified the proposed ODE by proposing alternative expressions for the injection term $Q(t)$ [see Wang et al. (2003), O'Brien and McPherron (2000)]. More recently Ballatore and Gonzalez (2014) have tried to generate empirical estimates for the injection and decay terms in Burton's equation.

Another important empirical model used to predict *Dst* is the *Nonlinear Auto-Regessive Moving Average with eXogenous inputs* (NARMAX) methodology developed in Billings et al. (1989), Balikhin et al. (2001), Zhu et al. (2006), Zhu et al. (2007), Boynton et al. (2011a), Boynton et al. (2011b) and Boynton et al. (2013). The NARMAX methodology builds models by constructing polynomial expansions of inputs and determines the best combinations of monomials to include in the refined model by using a criterion called the *error reduction ratio* (ERR). The parameters of the so called NARMAX OLS-ERR model are calculated by solving the *ordinary least squares* (OLS) problem arising from a quadratic objective function. The reader may refer to Billings (2013) for a detailed exposition of the NARMAX methodology.

Yet another family of forecasting methods is based on *Artificial Neural Networks* (ANN) that have been a popular choice for building predictive models. Researchers have employed both the standard *feed forward* and the more specialized *recurrent* architectures. Lundstedt et al. (2002) proposed an *Elman* recurrent network architecture called Lund *Dst*, which used the solar wind velocity, *interplanetary magnetic field* (IMF) and historical *Dst* data as inputs. Wing et al. (2005) used recurrent neural networks to predict *Kp*. Bala et al. (2009) originally proposed a *feed forward* network for predicting the K_p index which used the *Boyle coupling function* (Boyle et al., 1997). The same architecture is adapted for prediction of *Dst* in Bala et al. (2009), popularly known as the Rice *Dst* model. Pallocchia et al. (2006) proposed a *neural network* model called EDDA to predict *Dst* using only the IMF data.

In light of the extensive list of modeling techniques employed for prediction of the *Dst* index, model comparison and evaluation becomes a crucial step for advancing the research domain. Rastätter et al. (2013) compared several physics based (convection, kinetic and magneto hydrodynamic) and empirical prediction models such as NARMAX, Rice *Dst* on 4 storm events which

occurred between 2001 and 2006. Amata et al. (2008) compared the EDDA and the Lund *Dst* models over the 2003-2005 period. The most extensive model comparison in terms of storm events was probably conducted in Ji et al. (2012) which compared six *Dst* models (see table 1) on a list of 63 geomagnetic storm events of varying intensities which occurred between 1998 and 2006. In the comparison done in Ji et al. (2012), the model proposed in Temerin and Li (2002) gives the best performance on the test set considered.

In *Dst* prediction, a seemingly trivial yet highly informative prediction method is represented by the so called *Persistence* model, which uses the previous value of *Dst* as the prediction for the next time step ($\hat{Dst}(t) = Dst(t-1)$). Due to high correlation between *Dst* values one hour apart ($Dst(t)$, $Dst(t-1)$), the *Persistence* model gives excellent predictive performance on error metrics, despite its lack of sophistication. In essence it is a trivial hypothesis and hence it should always be used as a base line to compare the performance of any proposed *Dst* algorithm. Moreover, given its zero computational cost, a model should at least outperform the *Persistence* model in order to be a viable candidate for advancing the science of space weather prediction. An obvious critique to the Persistence model is that it is not really predictive, in the sense that it cannot forecast a storm, until it has already commenced. However, most of the literature has used global metrics, such as the Root Mean Square Error (RMSE, see section 5) to evaluate models. It is in this context that we argue that the Persistence model should be regarded as the first candidate to outperform.

In this paper, we propose *Gaussian Process* models for OSA prediction of *Dst*. We use the results of Ji et al. (2012) as a starting point and compare our proposed models with the methods evaluated in that paper, while using the performance of the *Persistence* model as a base line.

We use hourly resolution measurements extracted from NASA/GSFC's OMNI data set through OMNIWeb (King and Papitashvili, 2005). The rest of the paper is organized as follows. Section 2 gives an introduction to the *Gaussian Process* methodology for regression and some technical details about their application. In section 3 two different *Gaussian Process* auto-regressive models are proposed and subsequently in sections 5 and 6, benchmarked against the *Persistence* as well as the other models outlined in Ji et al. (2012).

2. Methodology: Gaussian Process

Gaussian Processes first appeared in machine learning research in Neal (1996), as the limiting case of Bayesian inference performed on neural networks with infinitely many neurons in the hidden layers. Although their inception in the machine learning community is recent, their origins can be traced back to the geo-statistics research community where they are known as *Kriging* methods (Krige (1951)). In pure mathematics area *Gaussian Processes* have been studied extensively and their existence was first proven by Kolmogorov's extension theorem (Tao (2011)). The reader is referred to Rasmussen and Williams (2005) for an in depth treatment of Gaussian Processes in machine learning.

Let us assume that we want to model a process in which a scalar quantity y is described by $y = f(\mathbf{x}) + \epsilon$ where $f(.) : \mathbb{R}^d \rightarrow \mathbb{R}$ is an unknown scalar function of a multidimensional input vector $\mathbf{x} \in \mathbb{R}^d$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is Gaussian distributed noise with variance σ^2.

A set of labeled data points $(\mathbf{x}_i, y_i); i = 1 \cdots N$ can be conveniently expressed by a $N \times d$ data matrix $\mathbf{X}$ and a $N \times 1$ response vector $\mathbf{y}$, as shown in equations (1) and (2).

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}_{n \times d} \tag{1}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{n \times 1} \tag{2}$$

Our task is to infer the values of the unknown function $f(.)$ based on the inputs $\mathbf{X}$ and the noisy observations $\mathbf{y}$. We now assume that the joint distribution of $f(\mathbf{x}_i), i = 1 \cdots N$ is a multivariate Gaussian as shown in equations (3), (4) and (5).

$$\mathbf{f} = \begin{pmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_N) \end{pmatrix} \tag{3}$$

$$\mathbf{f}|\mathbf{x}_1, \cdots, \mathbf{x}_N \sim \mathcal{N}(\mu, \mathbf{\Lambda}) \tag{4}$$

$$p(\mathbf{f} \mid \mathbf{x}_1, \cdots, \mathbf{x}_N) = \frac{1}{(2\pi)^{n/2} det(\mathbf{\Lambda})^{1/2}} exp\left(-\frac{1}{2}(\mathbf{f} - \mu)^T \mathbf{\Lambda}^{-1}(\mathbf{f} - \mu)\right) \tag{5}$$

Here $\mathbf{f}$ is a $N \times 1$ vector consisting of the values $f(\mathbf{x}_i), i = 1 \cdots N$. In equation (4), $\mathbf{f}|\mathbf{x}_1, \cdots, \mathbf{x}_N$ denotes the conditional distribution of $\mathbf{f}$ with respect to the input data (i.e., $\mathbf{X}$) and $\mathcal{N}(\mu, \mathbf{\Lambda})$ represents a multivariate Gaussian distribution with mean vector μ and covariance matrix $\mathbf{\Lambda}$. The probability density function of this distribution $p(\mathbf{f} \mid \mathbf{x}_1, \cdots, \mathbf{x}_N)$ is therefore given by equation (5).

From equation (5), one can observe that in order to uniquely define the distribution of the process, it is required to specify μ and $\mathbf{\Lambda}$. For this probability density to be valid, there are further requirements imposed on $\mathbf{\Lambda}$:

1. Symmetry: $\mathbf{\Lambda}_{ij} = \mathbf{\Lambda}_{ji} \; \forall i, j \in 1, \cdots, N$
2. Positive Semi-definiteness: $\mathbf{z}^T \mathbf{\Lambda} \mathbf{z} \geq 0 \; \forall \mathbf{z} \in \mathbb{R}^N$

Inspecting the individual elements of μ and $\mathbf{\Lambda}$, we realise that they take the following form.

$$\mu_i = \mathbb{E}[f(\mathbf{x}_i)] := m(\mathbf{x}_i) \tag{6}$$

$$\Lambda_{ij} = \mathbb{E}[(f(\mathbf{x}_i) - \mu_i)(f(\mathbf{x}_j) - \mu_j)] := K(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

Here $\mathbb{E}$ denotes the expectation (average). The elements of μ and $\mathbf{\Lambda}$ are expressed as functions $m(\mathbf{x}_i)$ and $K(\mathbf{x}_i, \mathbf{x}_j)$ of the inputs $\mathbf{x}_i, \; \mathbf{x}_j$. Specifying the functions $m(\mathbf{x})$ and $K(\mathbf{x}, \mathbf{x}')$ completely specifies each element of μ and $\mathbf{\Lambda}$ and subsequently the finite dimensional distribution of $\mathbf{f}|\mathbf{x}_1, \cdots, \mathbf{x}_N$. In

most practical applications of *Gaussian Processes* the mean function is often defined as $m(\mathbf{x}) = 0$, which is not unreasonable if the data is standardized to have zero mean. *Gaussian Processes* are represented in machine learning literature using the following notation:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')) \tag{8}$$

2.1. Inference and Predictions

Our aim is to infer the function $f(\mathbf{x})$ from the noisy training data and generate predictions $f(\mathbf{x}_i^*)$ for a set of test points $\mathbf{x}_i^* : \forall i \in 1, \cdots, M$. We define $\mathbf{X}^*$ as the test data matrix whose rows are formed by $\mathbf{x}_i^*$ as shown in equation (9).

$$\mathbf{X}_* = \begin{pmatrix} (\mathbf{x}_1^*)^T \\ (\mathbf{x}_2^*)^T \\ \vdots \\ (\mathbf{x}_M^*)^T \end{pmatrix}_{M \times d} \tag{9}$$

Using the multivariate Gaussian distribution in equation (5) we can construct the joint distribution of $f(\mathbf{x})$ over the training and test points. The vector of training and test outputs $\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix}$ is of dimension $(N + M) \times 1$ and is constructed by appending the test set predictions $\mathbf{f}_*$ to the observed noisy measurements $\mathbf{y}$.

$$\mathbf{f}_* = \begin{pmatrix} f(\mathbf{x}_1^*) \\ f(\mathbf{x}_2^*) \\ \vdots \\ f(\mathbf{x}_M^*) \end{pmatrix}_{M \times 1} \tag{10}$$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} | \ \mathbf{X}, \mathbf{X}_* \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right) \tag{11}$$

Since we have noisy measurements of f over the training data, we add the noise variance σ^2 to the variance of f as shown in (11). The block matrix components of the $(N + M) \times (N + M)$ covariance matrix have the following structure.

1. $\mathbf{I}$: The $n \times n$ identity matrix.
2. $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)], \ i, j \in 1, \cdots, n$: Kernel matrix constructed from all couples obtained from the training data.
3. $\mathbf{K}_* = [K(\mathbf{x}_i, \mathbf{x}_j^*)], \ i \in 1, \cdots, n; j \in 1, \cdots, m$: Cross kernel matrix constructed from all couples between training and test data points.
4. $\mathbf{K}_{**} = [K(\mathbf{x}_i^*, \mathbf{x}_j^*)], \ i, j \in 1, \cdots, m$: Kernel matrix constructed from all couples obtained from the test data.

With the multivariate normal distribution defined in equation (11), probabilistic predictions f_* can be generated by constructing the conditional distribution $\mathbf{f}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*$. Since the original distribution of $\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} | \ \mathbf{X}, \mathbf{X}_*$ is a multivariate Gaussian, conditioning on a subset of elements $\mathbf{y}$ yields another Gaussian distribution whose mean and covariance can be calculated exactly, as in equation (12) (see Rasmussen and Williams (2005)).

$$\mathbf{f}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \Sigma_*), \tag{12}$$

where

$$\bar{\mathbf{f}}_* = \mathbf{K}_*^T[\mathbf{K} + \sigma^2\mathbf{I}]^{-1}\mathbf{y} \tag{13}$$

$$\Sigma_* = \mathbf{K}_{**} - \mathbf{K}_*^T\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{K}_* \tag{14}$$

The practical implementation of *Gaussian Process* models requires the inversion of the training data kernel matrix $[\mathbf{K} + \sigma^2\mathbf{I}]^{-1}$ to calculate the parameters of the predictive distribution $\mathbf{f}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*$. The computational complexity of this inference is dominated by the linear problem in Eq. (13), which can be solved via Cholesky decomposition, with a time complexity of $O(N^3)$, where N is the number of data points.

The distribution of $\mathbf{f}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*$ is known in Bayesian analysis as the *Posterior Predictive Distribution*. This illustrates a key difference between *Gaussian Processes* and other regression models such as *Neural Networks*, *Linear Models* and *Support Vector Machines*: a *Gaussian Process* model does not generate point predictions for new data but outputs a predictive distribution for the quantity sought, thus allowing to construct error bars on the predictions. This property of Bayesian models such as *Gaussian Processes* makes them very appealing for Space Weather forecasting applications.

The central design issue in applying *Gaussian Process* models is the choice of the function $K(\mathbf{x}, \mathbf{x}')$. The same constraints that apply to $\mathbf{\Lambda}$ also apply to the function K. In machine learning, these symmetric positive definite functions of two variables are known as *kernels*. Kernel based methods are applied extensively in data analysis i.e. regression, clustering, classification, density estimation (see Scholkopf and Smola (2001), Hofmann et al. (2008)).

2.2. Kernel Functions

For the success of a *Gaussian Process* model an appropriate choice of kernel function is paramount. The symmetry and positive semi-definiteness of *Gaussian Process* kernels implies that they represent inner-products between some basis function representation of the data. The interested reader is suggested to refer to Berlinet and Thomas-Agnan (2004), Scholkopf and Smola (2001) and Hofmann et al. (2008) for a thorough treatment of kernel functions and the rich theory behind them. Some common kernel functions used in machine learning include the radial basis function (RBF) kernel $K(\mathbf{x}, \mathbf{y}) = \frac{1}{2}exp(-\|\mathbf{x} - \mathbf{y}\|^2/l^2)$ and the polynomial kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\intercal\mathbf{y} + b)^d$.

The quantities l in the RBF, and b and d in the polynomial kernel are known as *hyper-parameters*. Hyper-parameters give flexibility to particular kernel structure, for example $d = 1, 2, 3, \cdots$ in the polynomial kernel represents linear, quadratic, cubic and higher order polynomials respectively. The method of assigning values to the *hyper-parameters* is crucial in the model building process.

In this study, we construct Gaussian Process regression models with linear kernels as shown in equation (15), for *one step ahead* prediction of the *Dst* index. Our choice of kernel leaves us with only one adjustable hyper-parameter, the model noise σ.

$$K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\intercal \mathbf{y} \tag{15}$$

We initialize a grid of values for the model noise σ and use the error on a predefined validation data set to choose the best performing value of σ. While constructing the grid of possible values for σ it must be ensured that $\sigma > 0$ so that the kernel matrix constructed on the training data is non-singular.

3. One Step Ahead Prediction

Below in equations (16) - (19) we outline a *Gaussian Process* formulation for *OSA* prediction of *Dst*. A vector of features $\mathbf{x}_{t-1}$ is used as input to an unknown function $f(\mathbf{x}_{t-1})$.

The features $\mathbf{x}_{t-1}$ can be any collection of quantities in the hourly resolution OMNI data set. Generally $\mathbf{x}_{t-1}$ are time histories of *Dst* and other important variables such as plasma pressure $p(t)$, solar wind speed $V(t)$, z component of the interplanetary magnetic field $B_z(t)$.

$$\begin{aligned} Dst(t) &= f(\mathbf{x}_{t-1}) + \epsilon && (16)\\ \epsilon &\sim \mathcal{N}(0, \sigma^2) && (17)\\ f(x_t) &\sim \mathcal{GP}(m(\mathbf{x}_t), K(\mathbf{x}_t, \mathbf{x}_s)) && (18)\\ K(\mathbf{x}, \mathbf{y}) &= \mathbf{x}^\intercal \mathbf{y} + b && (19) \end{aligned}$$

In the following we consider two choices for the input features $\mathbf{x}_{t-1}$ leading to two variants of *Gaussian Process* regression for *Dst* time series prediction.

3.1. Gaussian Process Auto-Regressive (GP-AR)

The simplest auto-regressive models for *OSA* prediction of *Dst* are those that use only the history of *Dst* to construct input features for model training. The input features $\mathbf{x}_{t-1}$ at each time step are the history of $Dst(t)$ until a time lag of p hours.

$$\mathbf{x}_{t-1} = (Dst(t-1), \cdots, Dst(t-p+1))$$

3.2. Gaussian Process Auto-Regressive with eXogenous inputs (GP-ARX)

Auto-regressive models can be augmented by including exogenous quantities in the inputs $\mathbf{x}_{t-1}$ at each time step, in order to improve predictive accuracy. *Dst* gives a measure of ring currents, which are modulated by plasma sheet particle injections into the inner magnetosphere during sub-storms. Studies have shown that the substorm occurrence rate increases with solar wind velocity (high speed streams) (Kissinger et al., 2011; Newell et al., 2016). Prolonged southward interplanetary magnetic field (IMF) z-component (B_z) is needed for sub-storms to occur (McPherron et al., 1986).

An increase in the solar wind electric field, VB_z, can increase the dawn-dusk electric field in the magnetotail, which in turn determines the amount of plasma sheet particle that move to the inner magnetosphere (Friedel et al., 2001). Therefore, our exogenous parameters consist of solar wind velocity, IMF B_z, and VB_z.

In this model we choose distinct time lags p and p_{ex} for auto-regressive and exogenous variables respectively. In addition we explicitly include the product VB_z which is a proxy for the electric field, as an input.

$$\begin{aligned}\mathbf{x}_{t-1} &= (Dst(t-1), \cdots, Dst(t-p+1),\\ &\quad V(t-1), \cdots, V(t-p_{ex}+1),\\ &\quad B_z(t-1), \cdots, B_z(t-p_{ex}+1),\\ &\quad VB_z(t-1), \cdots, VB_z(t-p_{ex}+1))\end{aligned}$$

4. Model Training and Validation

Before running performance bench marks for *OSA Dst* prediction on the storm events in Ji et al. (2012), training and model selection of *GP-AR* and *GP-ARX* models on independent data sets must be performed. For this purpose we choose segments 00:00 1 January 2008 - 10:00 11 January 2008 for training and 00:00 15 November 2014 - 23:00 1 December 2014 for model selection.

Although the training and model selection data sets both do not have a geomagnetic storm in them, this would not degrade the performance of *GP-AR* and *GP-ARX* because the linear polynomial kernel describes a non stationary and self similar Gaussian Process. This implies that for two events where the time histories of Dst, V and B_z are not close to each other but can be expressed as a diagonal rescaling of time histories observed in the training data, the predictive distribution is a linearly rescaled version of the training data Dst distribution.

The computational complexity of calculation of the predictive distribution is $O(N^3)$, as discussed in section 2.1. This can limit the size of the covariance matrix constructed from the training data. Note that this computation overhead is paid for every unique assignment to the model hyper-parameters. However, our chosen training set has a size of 250 which is still very much below the computational limits of the method and in our case, it has been decided by studying the performance of the models for increasing training sets. We have noticed that using training sets with more than 250 values does not increase the overall performance in *OSA* prediction of Dst.

Model selection of *GP-AR* and *GP-ARX* is performed by grid search methodology and the root mean square error (*RMSE*) on the validation set is used to select the values of the model hyper-parameters. In the experiments performed the best performing values of the hyper-parameters outputted by the grid search routine are $\sigma = 0.45$ for the *GP-ARX* and $\sigma = 1.173$ for *GP-AR*.

The time lags chosen for *GP-AR* are $p = 6$ while for *GP-ARX* they are $p = 6$, $p_{ex} = 1$. These values are arrived at by building *GP-AR* and *GP-ARX* models for different values of p and choosing the model yielding the best performance on a set of storm events during the periods $1995 - 1997$ and $2011 - 2014$.

Box plots 1 and 2 show the performance of *GP-ARX* models having values of $p = 3, 4, 5, 6, 7$. We observe that the median *RMSE* and *CC* degrade slightly from $p = 3$ to $p = 7$ and models with $p = 6$

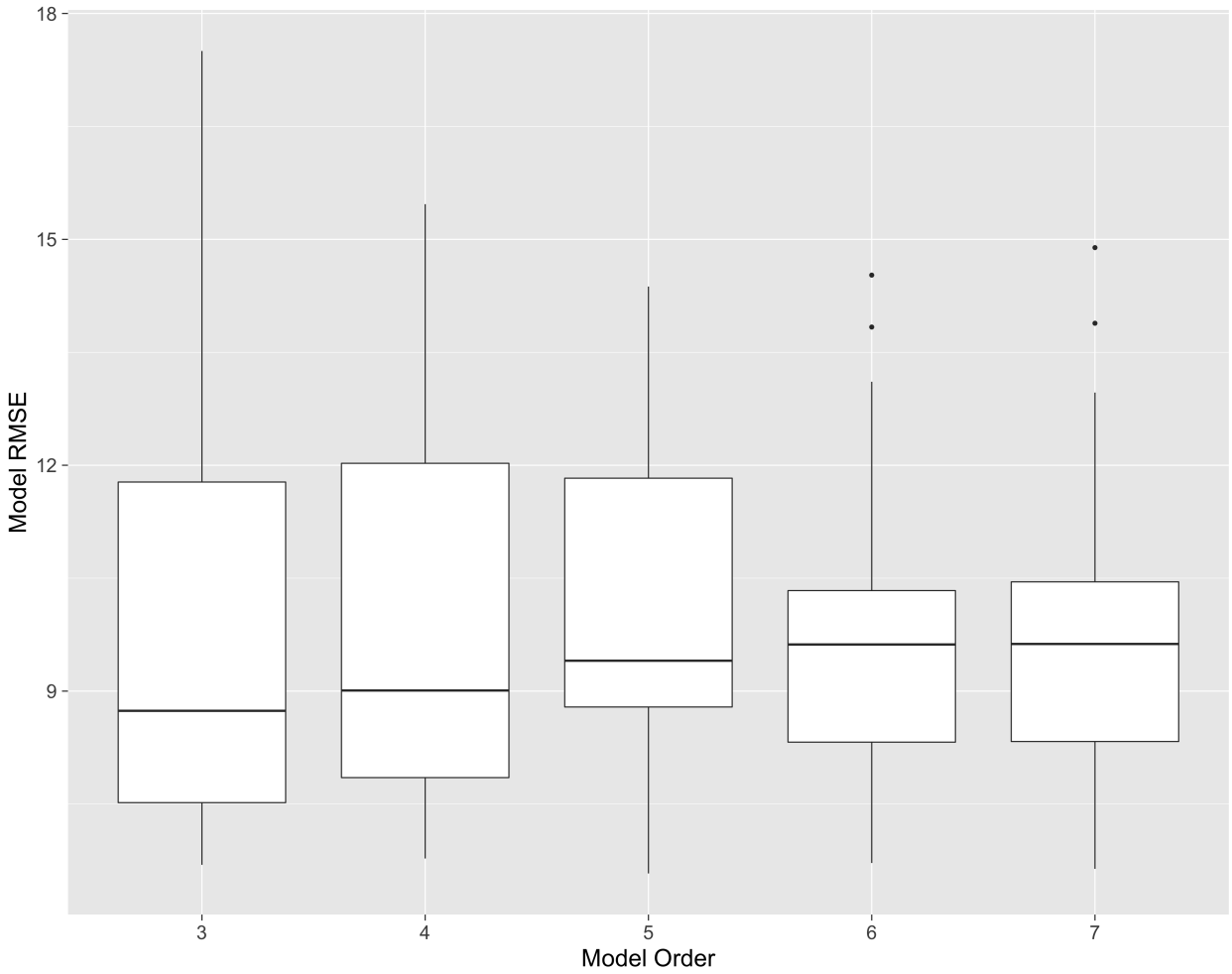

Fig. 1. Averaged RMSE performance of *GP-ARX* models for different values of auto-regressive order p

have a much smaller variance in performance making them more robust to outliers. The choice of $p = 6$ thus gives a good trade off between model robustness and complexity.

This approach for choosing the time lags does not lead to bias towards models which only perform well on the data set of Ji et al. (2012) because the storms used to choose the model order have no overlap with it. Moreover, Gaussian Processes are not easily susceptible to overfitting when the number of model hyper-parameters is small.

5. Experiments

Ji et al. (2012) compare six *one step ahead Dst* models as listed in table 1 in their paper by compiling a list of 63 geomagnetic storm events of varying strengths which occurred in the period 1998-2006. This serves as an important stepping stone for a systematic comparison of existing and new prediction techniques because performance metrics are averaged over a large number of storm

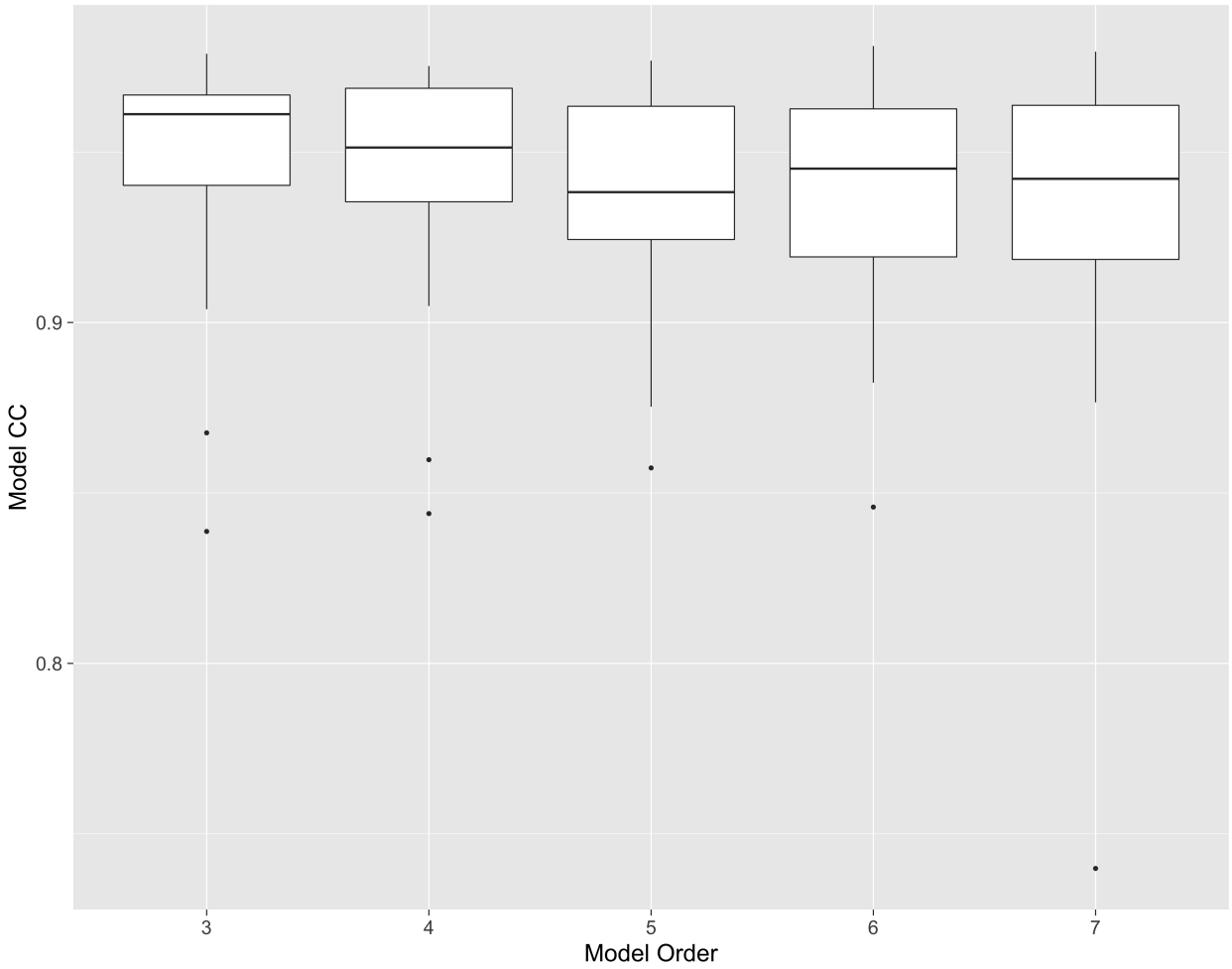

Fig. 2. Averaged Correlation Coefficient of *GP-ARX* models for different values of auto-regressive order p

events that occurred over a 8 year period. They compare the averaged performance of these models on four key performance metrics.

1. The root mean square error.

$$RMSE = \sqrt{\sum_{t=1}^{n}(Dst(t) - \hat{D}st(t))^2/n} \tag{20}$$

2. Correlation coefficient between the predicted and actual value of Dst.

$$CC = Cov(Dst, \hat{D}st)/\sqrt{Var(Dst)Var(\hat{D}st)} \tag{21}$$

3. The error in prediction of the peak negative value of Dst for a storm event.

$$\Delta Dst_{min} = Dst(t^*) - \hat{D}st(t^{**}),\ t^* = argmin(Dst(t)),\ t^{**} = argmin(\hat{D}st(t)) \tag{22}$$

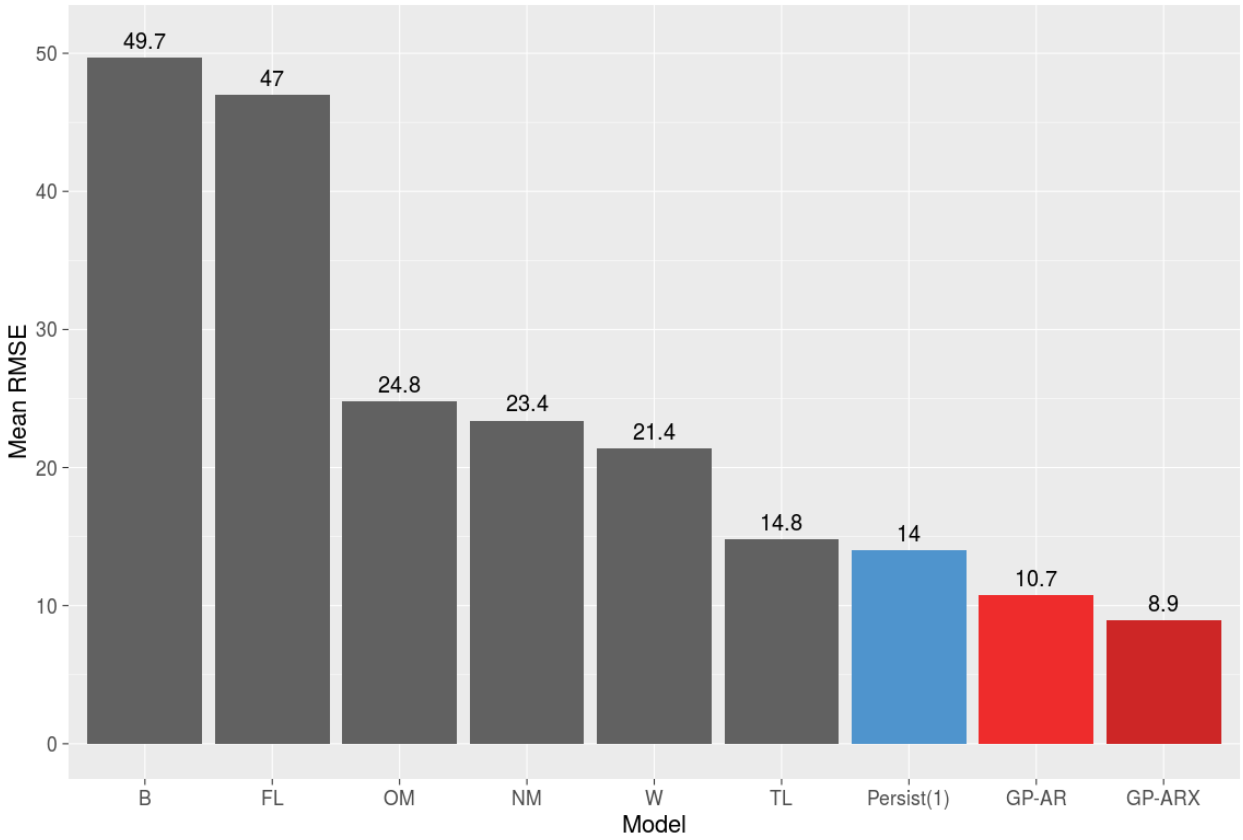

Fig. 3. Averaged RMSE performance of *GP-AR*, *GP-ARX* and the *Persistence* model versus results in Ji et al. (2012).

4. The error in predicting the timing of a storm peak, also called *timing error* $|\Delta t_{peak}|$.

$$\Delta t_{peak} = argmin_t(D_{st}(t)) - argmin_t(\hat{D}_{st}(t)), \quad t \in \{1, \cdots, n\} \tag{23}$$

Apart from the metrics above, we also generate the empirical distribution of relative errors $|\frac{Dst - \hat{Dst}}{Dst}|$ for the *GP-AR*, *GP-ARX*, *NM* and *TL* models.

5.1. Model Comparison

Table 1 gives a brief enumeration of the models compared. We use the same performance benchmark results published in Ji et al. (2012) and we add the results obtained from testing *GP-AR*, *GP-ARX* as well as the *Persistence* model on the same list of storms.

For the purpose of generating the cumulative probability plots of the model errors, we need to obtain hourly predictions for all the storm events using the *NARMAX* and *TL* models. In the case of the *NM* model, the formula outlined in Boynton et al. (2011a) is used to generate hourly predictions. For the *TL* model we use real time predictions listed on their website (`http://lasp.colorado.edu/home/spaceweather/`) corresponding to all the storm events. The real time predictions are at a frequency of ten minutes and hourly predictions are generated by averaging the *Dst* predictions for each hour.

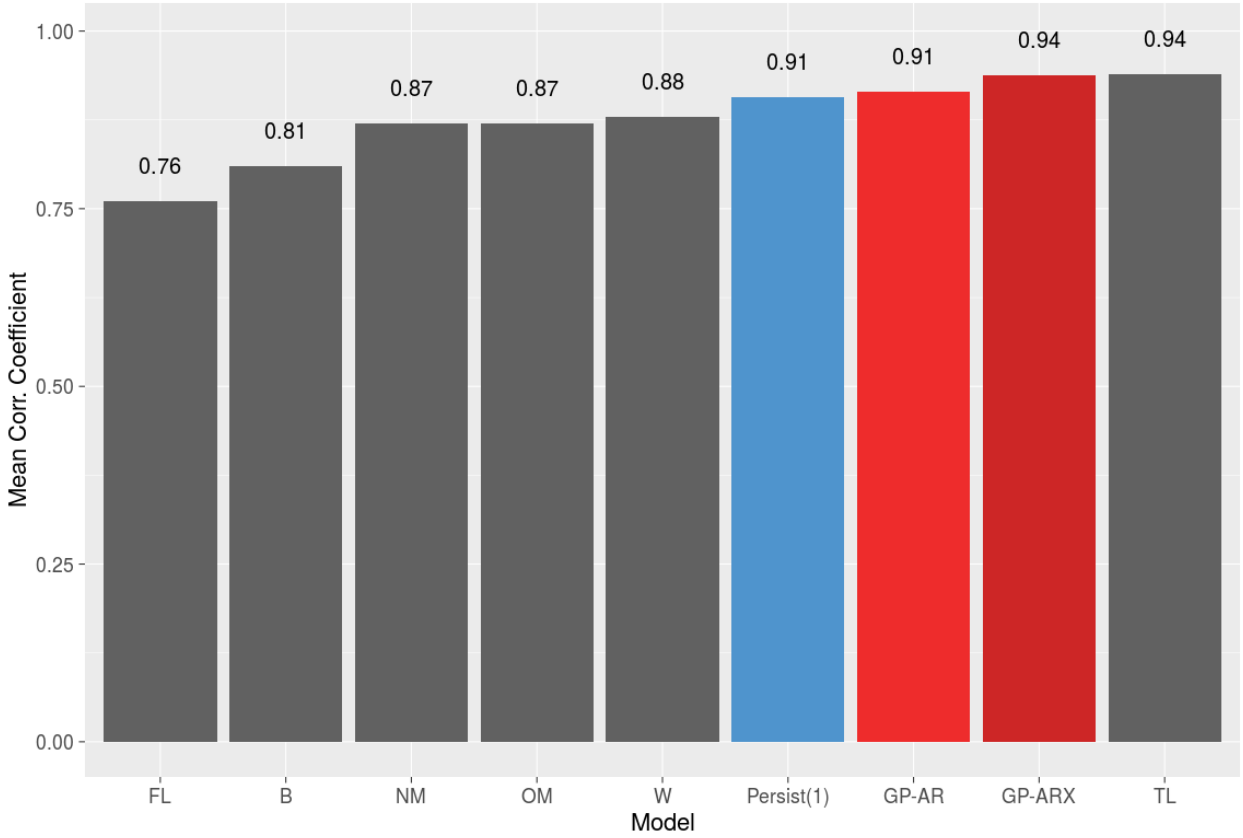

Fig. 4. Averaged cross correlation performance of *GP-AR*, *GP-ARX* and the *Persistence* model versus results in Ji et al. (2012).

Table 1. One Step Ahead *Dst* prediction models compared in Ji et al. (2012)

Model	Reference	Description
TL	Temerin and Li (2002)	Auto regressive model decomposable into additive terms.
NM	Boynton et al. (2011a)	Non linear auto regressive with exogenous inputs.
B	Burton et al. (1975)	Prediction of *Dst* by solving ODE having injection and decay terms.
W	Wang et al. (2003)	Obtained by modification of injection term in Burton et al. (1975).
FL	Fenrich and Luhmann (1998)	Uses polarity of magnetic clouds to predict geomagnetic response.
OM	O'Brien and McPherron (2000)	Modification of injection term in Burton et al. (1975).

With respect to the *TL* model one important caveat must be noted. The training of the *TL* model was performed on data from 1996-2002 which overlaps with a large number of the storm events, therefore the experimental procedure has a strong bias towards *TL* model.

6. Results

Figures 3, 4, 5 and 6 compare the performance of *GP-AR*, *GP-ARX* and the *Persistence* model with the existing results of Ji et al. (2012). Figure 3 compares the average RMSE of the predictions over

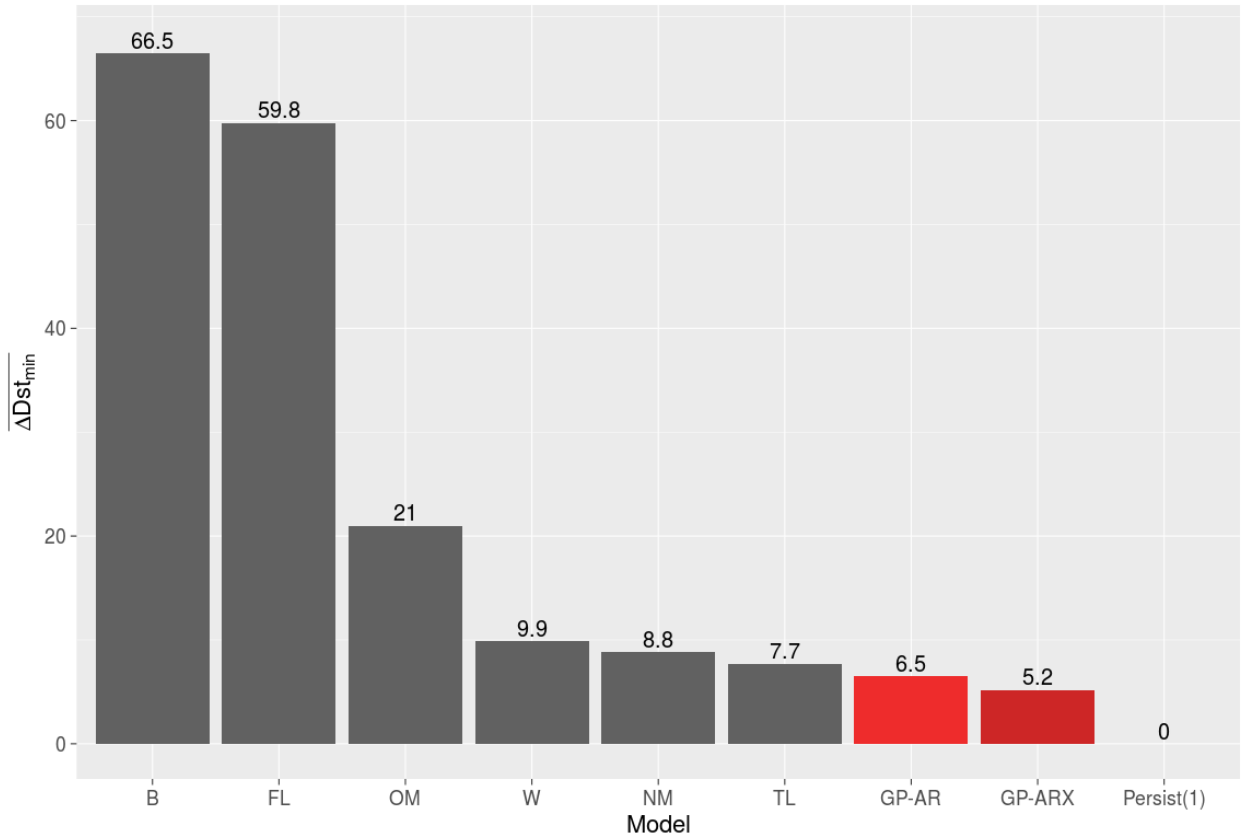

Fig. 5. Averaged ΔDst_{min} of *GP-AR*, *GP-ARX* and the *Persistence* model versus results in Ji et al. (2012).

the list of 63 storms. One can see that the *GP-ARX* model gives an improvement of approximately 40% in RMSE with respect to the *TL* model and 62% improvement with respect to the *NM* model. It can also be seen that the *Persistence* model gives better *RMSE* performance than the models compared in Ji et al. (2012).

Figure 4 shows the comparison of correlation coefficients between model predictions and actual *Dst* values. From the results of Ji et al. (2012), the *TL* model gives the highest correlation coefficient of 94%, but that is not surprising considering the fact that there is a large overlap between the training data used to train it and storm events. Even so the *GP-ARX* model gives a comparable correlation coefficient to *TL*, although the training set of the *GP-ARX* model has no overlap with the storm events, which is not the case for *TL*.

Figure 5 compares the different models with respect to accuracy in predicting the peak value of *Dst* during storm events (ΔDst_{min} averaged over all storms). In the context of this metric the *GP-ARX* model gives a 32% improvement over the *TL* model and 41% over the *NM* model. It is trivial to note that by definition, the *Persistence* model ($\hat{Dst}(t) = Dst(t-1)$) will have $\Delta Dst_{min} = 0$ for every storm event.

Figure 6 compares the time discrepancy of predicting the storm peak. By definition, the *Persistence* model will always have $\Delta t_{peak} = -1$. The *GP-AR* and *GP-ARX* models give better performance than the other models in terms of timing error in prediction of storm peaks.

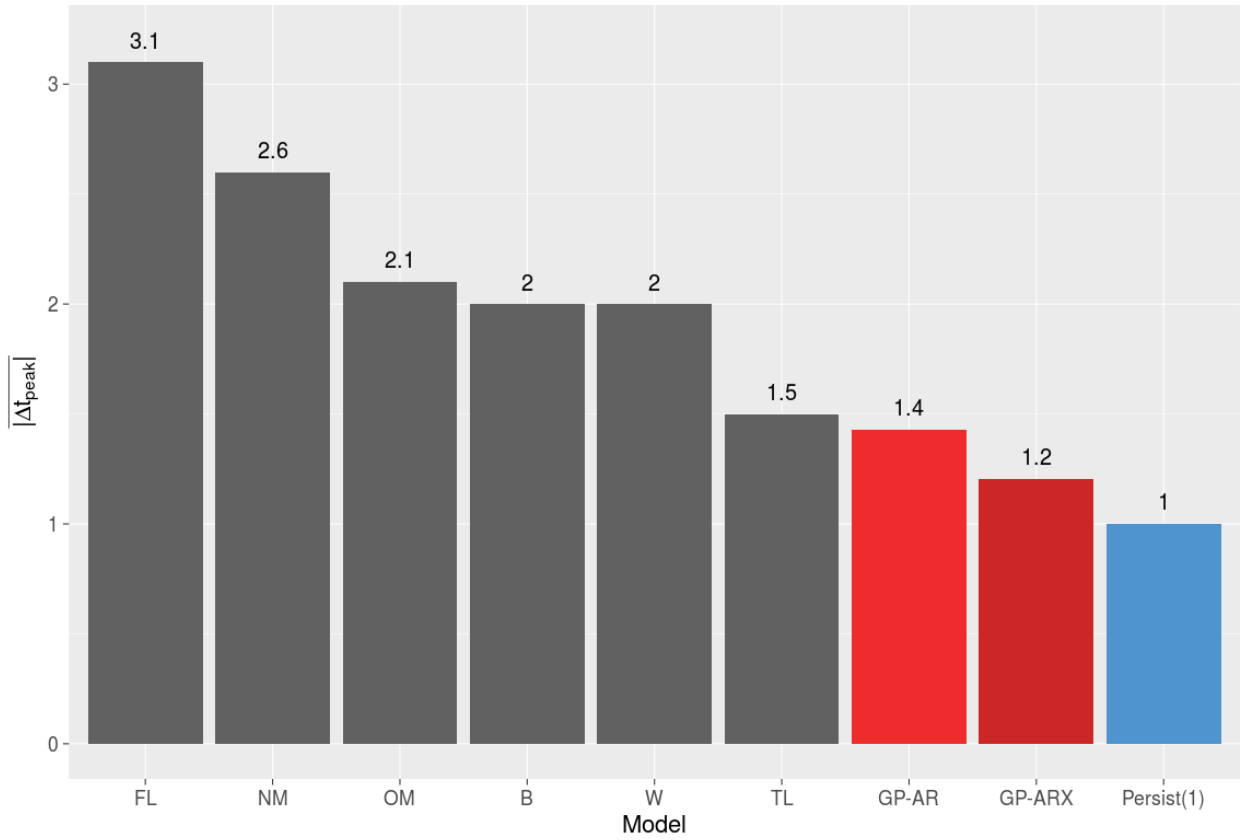

Fig. 6. Averaged absolute timing error of *GP-AR*, *GP-ARX* and the *Persistence* model versus results in Ji et al. (2012).

Figure 7 compares the empirical cumulative distribution functions of the absolute value of the relative errors across all storm events in the experiments. One can see that 95% of the hourly predictions have a relative error smaller than 20% for *GP-ARX*, 50% for *NM* and 60% for *TL*.

In Figure 8 we show the OSA predictions generated by the *TL*, *NM* and *GP-ARX* models for one event from the list of storms in Ji et al. (2012). This storm is interesting due to its strength ($Dst_{min} = -289\ nT$) and the presence of two distinct peaks. In this particular event the *GP-ARX* model approximates the twin peaks as well as the onset and decay phases quite faithfully, contrary the *NM* and *TL* model predictions. The *TL* model recognises the presence of twin peaks but overestimates the second one as well as the decay phase, while the *NM* model is much delayed in the prediction of the initial peak and fails to approximate the time and intensity of the second peak.

As noted in section 2.1, *Gaussian Process* models generate probabilistic predictions instead of point predictions, giving the modeller the ability to specify error bars on the predictions. In Figure 9, *OSA* predictions along with lower and upper error bars generated by the *GP-ARX* model are depicted for the event shown in Figure 8. The error bars are calculated at one standard deviation on either side of the mean prediction, using the posterior predictive distribution in equations (13) and (14). Since the predictive distributions are Gaussian, one standard deviation on either side of the mean corresponds to a confidence interval of 68%.

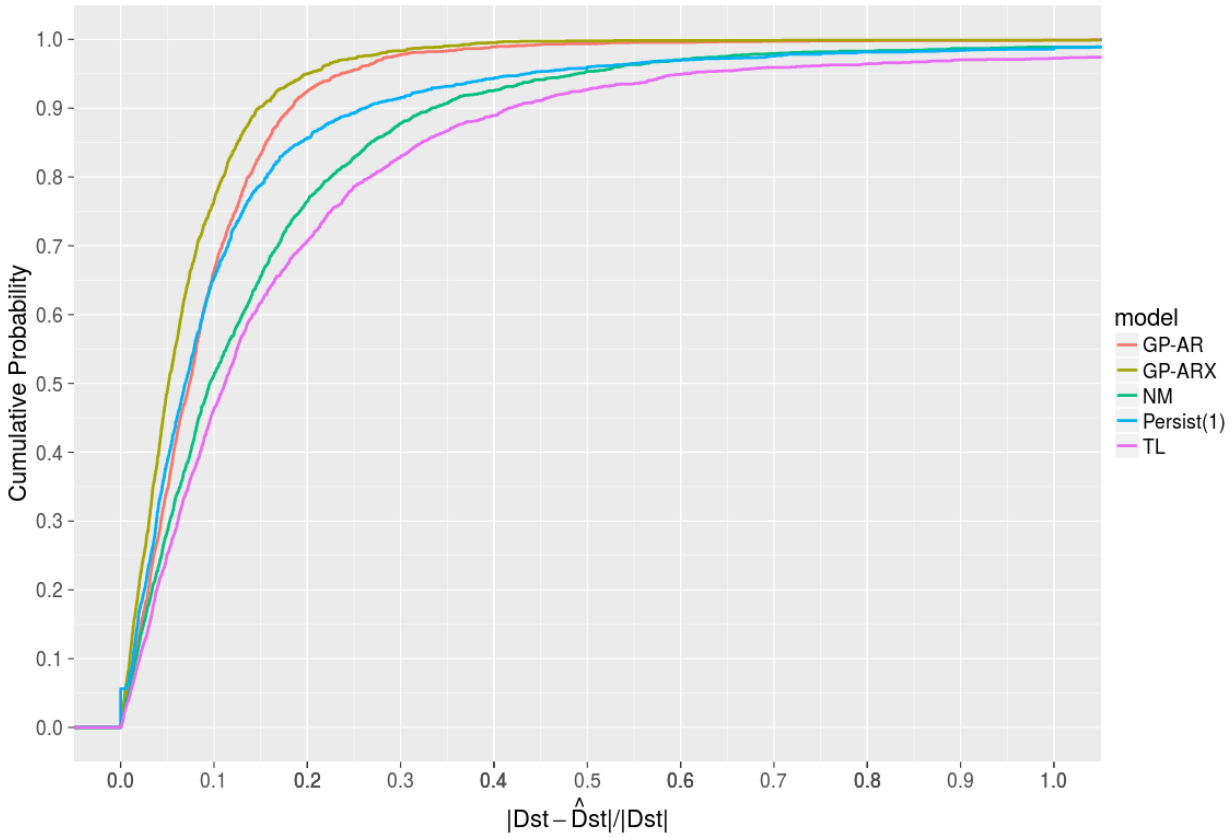

Fig. 7. Comparison of empirical distribution of relative errors for *GP-AR*, *GP-ARX*, *NM* and *TL* on the test set of Ji et al. (2012).

7. Conclusions

In this paper, we proposed two *Gaussian Process* auto-regressive models, *GP-ARX* and *GP-AR*, to generate hourly predictions of *Dst*. We compared the performance of our proposed models with the *Persistence* model and six existing model benchmarks reported in Ji et al. (2012). Predictive performance was compared on *RMSE*, *CC*, ΔDst_{min} and $|\Delta t_{peak}|$ calculated on a list of 63 geomagnetic storms from Ji et al. (2012).

Our results can be summarized as follows.

1. *Persistence* model must be central in the model evaluation process in the context of *one step ahead* prediction of the *Dst* index. It is clear that the *persistence* behavior in the *Dst* values is very strong i.e. the trivial predictive model $\hat{Dst}(t) = Dst(t-1)$ gives excellent performance according to the metrics chosen. In fact it can be seen in Figure 3 that the *Persistence* model outperforms every model compared by Ji et al. (2012) in the context of OSA prediction of *Dst*. Therefore any model that is proposed to tackle the OSA prediction problem for *Dst* should be compared to the *Persistence* model and show visible gains above it.

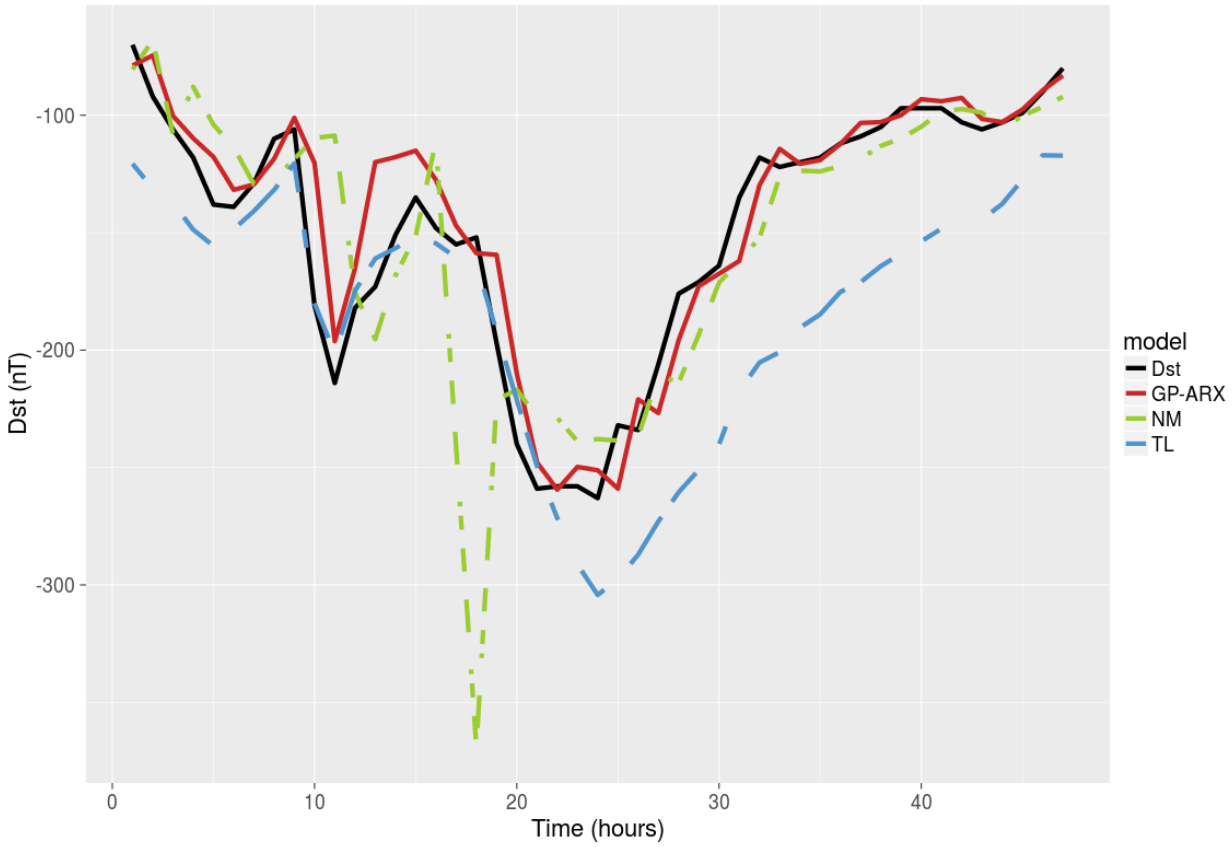

Fig. 8. Comparison of OSA predictions generated by the *NM*, *TL* and *GP-ARX* models for the storm event 9^{th} November 2004 11:00 UTC - 11^{th} November 2004 09:00 UTC

2. *Gaussian Process* AR and ARX models give encouraging benefits in OSA prediction even when compared to the *Persistence* model. Leveraging the strengths of the Bayesian approach, they are able to learn robust predictors from data.

   If one compares the training sets used by all the models, one can appreciate that the models presented here need relatively small training and validations sets: the training set contains 250 instances, while the validation set contains 432 instances. On the other hand the *NM* model uses one year (8760 instances) of data to learn the formula outlined in Boynton et al. (2011a) and the *TL* model uses six years of data (52560 instances) for training.

   The encouraging results of Gaussian Processes illustrate the strengths of the Bayesian approach to predictive modeling. Since the GP models generate predictive distributions for test data and not just point predictions they lend themselves to the requirements of space weather prediction very well because of the need to generate error bars on predictions.

Acknowledgements. We acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb (or CDAWeb or ftp) service, and OMNI data. We also acknowledge the authors of Temerin and Li (2002) for their scientific inputs towards understanding the *TL* model and assistance in access of its generated predictions. Simon Wing acknowledges supports from CWI and NSF Grant AGS-1058456 and NASA Grants (NNX13AE12G, NNX15AJ01G, NNX16AC39G).

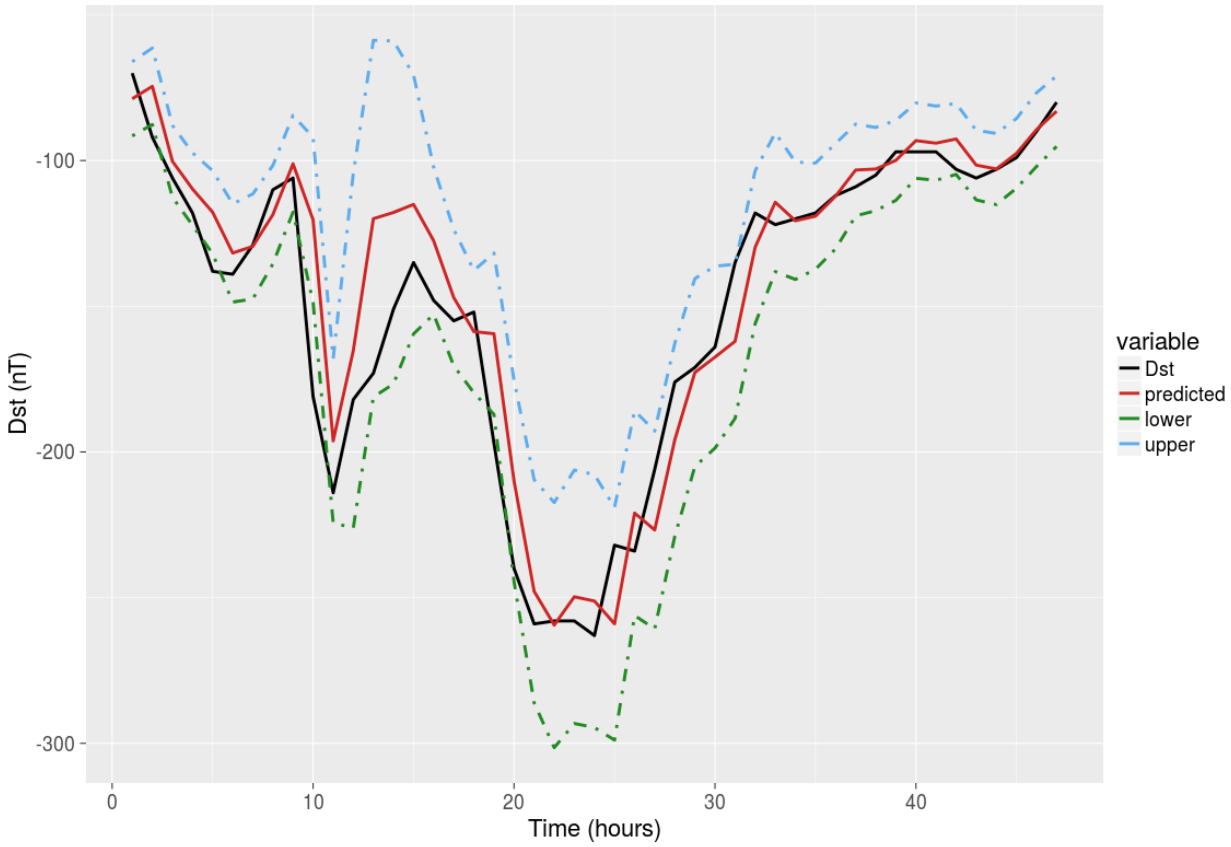

Fig. 9. OSA predictions with error bars ±1 standard deviation, generated by the *GP-ARX* model for the storm event 9^{th} November 2004 11:00 UTC - 11^{th} November 2004 09:00 UTC

References

Amata, E., G. Pallocchia, G. Consolini, M. Marcucci, and I. Bertello. Comparison between three algorithms for Dst predictions over the 20032005 period. *Journal of Atmospheric and Solar-Terrestrial Physics*, **70**(2-4), 496–502, 2008. 10.1016/j.jastp.2007.08.041, URL http://linkinghub.elsevier.com/retrieve/pii/S1364682607003082. 1

Bala, R., P. H. Reiff, and J. E. Landivar. Real-time prediction of magnetospheric activity using the Boyle Index. *Space Weather*, **7**(4), n/a–n/a, 2009. 10.1029/2008SW000407, URL http://dx.doi.org/10.1029/2008SW000407. 1

Balikhin, M. A., O. M. Boaghe, S. A. Billings, and H. S. C. K. Alleyne. Terrestrial magnetosphere as a nonlinear resonator. *Geophysical Research Letters*, **28**(6), 1123–1126, 2001. 10.1029/2000GL000112, URL http://dx.doi.org/10.1029/2000GL000112. 1

Ballatore, P., and W. D. Gonzalez. On the estimates of the ring current injection and decay. *Earth, Planets and Space*, **55**(7), 427–435, 2014. 10.1186/BF03351776, URL http://dx.doi.org/10.1186/BF03351776. 1

Bartels, J., and J. Veldkamp. International data on magnetic disturbances, second quarter, 1949. *Journal of Geophysical Research*, **54**(4), 399–400, 1949. 10.1029/JZ054i004p00399, URL http://dx.doi.org/10.1029/JZ054i004p00399. 1

Berlinet, A., and C. Thomas-Agnan. Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer US, 2004. ISBN 978-1-4613-4792-7. 10.1007/978-1-4419-9096-9, URL http://www.springer.com/us/book/9781402076794. 2.2

Billings, S. A. Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains. John Wiley & Sons, 2013. 1

Billings, S. A., S. Chen, and M. J. Korenberg. Identification of MIMO non-linear systems using a forward-regression orthogonal estimator. *International Journal of Control*, **49**(6), 2157–2189, 1989. 10.1080/00207178908559767, http://www.tandfonline.com/doi/pdf/10.1080/00207178908559767, URL http://www.tandfonline.com/doi/abs/10.1080/00207178908559767. 1

Boyle, C., P. Reiff, and M. Hairston. Empirical polar cap potentials. *Journal of Geophysical Research*, **102**(A1), 111–125, 1997. 1

Boynton, R. J., M. A. Balikhin, S. A. Billings, G. D. Reeves, N. Ganushkina, M. Gedalin, O. A. Amariutei, J. E. Borovsky, and S. N. Walker. The analysis of electron fluxes at geosynchronous orbit employing a NARMAX approach. *Journal of Geophysical Research: Space Physics*, **118**(4), 1500–1513, 2013. 10.1002/jgra.50192, URL http://dx.doi.org/10.1002/jgra.50192. 1

Boynton, R. J., M. A. Balikhin, S. A. Billings, A. S. Sharma, and O. A. Amariutei. Data derived NARMAX Dst model. *Annales Geophysicae*, **29**(6), 965–971, 2011a. 10.5194/angeo-29-965-2011, URL http://www.ann-geophys.net/29/965/2011/. 1, 5.1, 1, 2

Boynton, R. J., M. A. Balikhin, S. A. Billings, H. L. Wei, and N. Ganushkina. Using the NARMAX OLS-ERR algorithm to obtain the most influential coupling functions that affect the evolution of the magnetosphere. *Journal of Geophysical Research: Space Physics*, **116**(A5), n/a–n/a, 2011b. 10.1029/2010JA015505, URL http://dx.doi.org/10.1029/2010JA015505. 1

Burton, R. K., R. L. McPherron, and C. T. Russell. An empirical relationship between interplanetary conditions and Dst. *Journal of Geophysical Research*, **80**(31), 4204–4214, 1975. 10.1029/JA080i031p04204, URL http://dx.doi.org/10.1029/JA080i031p04204. 1, 1

Davis, T. N., and M. Sugiura. Auroral electrojet activity index AE and its universal time variations. *Journal of Geophysical Research*, **71**(3), 785–801, 1966. 10.1029/JZ071i003p00785, URL http://dx.doi.org/10.1029/JZ071i003p00785. 2

Dessler, A. J., and E. N. Parker. Hydromagnetic theory of geomagnetic storms. *Journal of Geophysical Research*, **64**(12), 2239–2252, 1959. 10.1029/JZ064i012p02239, URL http://dx.doi.org/10.1029/JZ064i012p02239. 3

Fenrich, F. R., and J. G. Luhmann. Geomagnetic response to magnetic clouds of different polarity. *Geophysical Research Letters*, **25**(15), 2999–3002, 1998. 10.1029/98GL51180, URL http://dx.doi.org/10.1029/98GL51180. 1

Friedel, R. H. W., H. Korth, M. G. Henderson, M. F. Thomsen, and J. D. Scudder. Plasma sheet access to the inner magnetosphere. *Journal of Geophysical Research: Space Physics*, **106**(A4), 5845–5858, 2001. 10.1029/2000JA003011, URL http://dx.doi.org/10.1029/2000JA003011. 3.2

Hofmann, T., B. Schlkopf, and A. J. Smola. Kernel methods in machine learning. *Ann. Statist.*, **36**(3), 1171–1220, 2008. 10.1214/009053607000000677, URL http://dx.doi.org/10.1214/009053607000000677. 2.1, 2.2

Ji, E. Y., Y. J. Moon, N. Gopalswamy, and D. H. Lee. Comparison of Dst forecast models for intense geomagnetic storms. *Journal of Geophysical Research: Space Physics*, **117**(3), 1–9, 2012. 10.1029/2011JA016872. 1, 4, 4, 5, 3, 5.1, 4, 1, 6, 5, 6, 6, 7, 7, 1

King, J. H., and N. E. Papitashvili. Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, **110**(A2), n/a—-n/a, 2005. 10.1029/2004JA010649, URL http://dx.doi.org/10.1029/2004JA010649. 1

Kissinger, J., R. L. McPherron, T.-S. Hsu, and V. Angelopoulos. Steady magnetospheric convection and stream interfaces: Relationship over a solar cycle. *Journal of Geophysical Research: Space Physics*, **116**(A5), n/a–n/a, 2011. A00I19, 10.1029/2010JA015763, URL http://dx.doi.org/10.1029/2010JA015763. 3.2

Krige, d. g. A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand. publisher not identified, 1951. URL https://books.google.co.in/books?id=M6jASgAACAAJ. 2

Lundstedt, H., H. Gleisner, and P. Wintoft. Operational forecasts of the geomagnetic Dst index. *Geophysical Research Letters*, **29**(24), 34–1–34–4, 2002. 2181, 10.1029/2002GL016151, URL http://dx.doi.org/10.1029/2002GL016151. 1

McPherron, R. L., T. Terasawa, and A. Nishida. Solar Wind Triggering of Substorm Expansion Onset. *Journal of geomagnetism and geoelectricity*, **38**(11), 1089–1108, 1986. 10.5636/jgg.38.1089. 3.2

Neal, R. M. Bayesian Learning for Neural Networks. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. ISBN 0387947248. 2

Newell, P., K. Liou, J. Gjerloev, T. Sotirelis, S. Wing, and E. Mitchell. Substorm probabilities are best predicted from solar wind speed. *Journal of Atmospheric and Solar-Terrestrial Physics*, **146**, 28 – 37, 2016. Http://dx.doi.org/10.1016/j.jastp.2016.04.019, URL http://www.sciencedirect.com/science/article/pii/S1364682616301195. 3.2

O'Brien, T. P., and R. L. McPherron. An empirical phase space analysis of ring current dynamics: Solar wind control of injection and decay. *Journal of Geophysical Research: Space Physics*, **105**(A4), 7707–7719, 2000. 10.1029/1998JA000437, URL http://dx.doi.org/10.1029/1998JA000437. 1, 1

Pallocchia, G., E. Amata, G. Consolini, M. F. Marcucci, and I. Bertello. Geomagnetic Dst index forecast based on IMF data only. *Annales Geophysicae*, **24**(3), 989–999, 2006. URL https://hal.archives-ouvertes.fr/hal-00318011. 1

Rasmussen, C. E., and C. K. I. Williams. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2005. ISBN 026218253X. 2, 2.1

Rastätter, L., M. M. Kuznetsova, A. Glocer, D. Welling, X. Meng, et al. Geospace environment modeling 2008-2009 challenge: Dst index. *Space Weather*, **11**(4), 187–205, 2013. 10.1002/swe.20036, URL http://doi.wiley.com/10.1002/swe.20036. 1

Scholkopf, B., and A. J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759. 2.1, 2.2

Tao, T. An Introduction to Measure Theory. Graduate studies in mathematics. American Mathematical Society, 2011. ISBN 9780821869192. URL https://books.google.nl/books?id=HoGDAwAAQBAJ. 2

Temerin, M., and X. Li. A new model for the prediction of Dst on the basis of the solar wind. *Journal of Geophysical Research: Space Physics*, **107**(A12), SMP 31–1–SMP 31–8, 2002. 10.1029/2001JA007532, URL http://dx.doi.org/10.1029/2001JA007532. 1, 1, 7

Wang, C. B., J. K. Chao, and C.-H. Lin. Influence of the solar wind dynamic pressure on the decay and injection of the ring current. *Journal of Geophysical Research: Space Physics*, **108**(A9), n/a–n/a, 2003. 1341, 10.1029/2003JA009851, URL http://dx.doi.org/10.1029/2003JA009851. 1, 1

Wing, S., J. R. Johnson, J. Jen, C.-I. Meng, D. G. Sibeck, K. Bechtold, J. Freeman, K. Costello, M. Balikhin, and K. Takahashi. Kp forecast models. *Journal of Geophysical Research: Space Physics*, **110**(A4), n/a–n/a, 2005. A04203, 10.1029/2004JA010500, URL http://dx.doi.org/10.1029/2004JA010500. 1

Zhu, D., S. A. Billings, M. Balikhin, S. Wing, and D. Coca. Data derived continuous time model for the Dst dynamics. *Geophysical Research Letters*, **33**(4), n/a–n/a, 2006. 10.1029/2005GL025022, URL http://dx.doi.org/10.1029/2005GL025022. 1

Zhu, D., S. A. Billings, M. A. Balikhin, S. Wing, and H. Alleyne. Multi-input data derived Dst model. *Journal of Geophysical Research: Space Physics*, **112**(A6), n/a–n/a, 2007. 10.1029/2006JA012079, URL http://dx.doi.org/10.1029/2006JA012079. 1