Probabilistic Forecasting of the Disturbance Storm Time Index: An Autoregressive Gaussian Process approach

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

Corresponding author: M. H. Chandorkar, Multiscale Dynamics, Centrum Wiskunde Informatica, Science Park 123, 1098XG Amsterdam, Netherlands. (mandar.chandorkar@cwi.nl)

[1]Multiscale Dynamics, Centrum Wiskunde Informatica (CWI), Amsterdam, 1098XG Amsterdam

[2]The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland, 20723, USA

Abstract. We present a methodology for generating probabilistic predictions for the *Disturbance Storm Time* (*Dst*) geomagnetic activity index. We focus on the *One Step Ahead* (OSA) prediction task and use the OMNI hourly resolution data to build our models.

Our proposed methodology is based on the technique of *Gaussian Process Regression* (GPR). Within this framework we develop two models; *Gaussian Process Auto-Regressive* (GP-AR) and *Gaussian Process Auto-Regressive with eXogenous inputs* (GP-ARX).

We also propose a criterion to aid model selection with respect to the order of auto-regressive inputs. Finally we test the performance of the GP-AR and GP-ARX models on a set of 63 geomagnetic storms between 1998 and 2006 and illustrate sample predictions with error bars for some of these events.

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, usually magnetometers, 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 activity 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 gives a measure of the weakening or strengthening of the Earth's equatorial magnetic field due to the weakening or strengthening of the ring currents and the 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.

Although much research has been done on prediction of the *Dst* index, much less has been done on probabilistic forecasting of *Dst*. One such work described in *McPherron et al.* [2013] involves identification of high speed solar wind streams using the WSA model (see *Wang and Sheeley* [1990]), using predictions of high speed streams to construct ensembles of *Dst* trajectories which yield the quartiles of *Dst* time series.

In this work we propose a technique for probabilistic forecasting of *Dst*, which yields a predictive distribution as a closed form expression. Our models take as input past values of *Dst*, solar wind speed and the z component of the *Interplanetary Magnetic Field* (IMF) and output a Gaussian distribution with a specific mean and variance as the OSA prediction of the *Dst*.

We use the *Gaussian Process Regression* methodology to construct auto-regressive models for *Dst* and show how to perform exact inference in this framework. We further outline a methodology to perform model selection with respect to its free parameters and time histories.

The remainder of this paper is organised as follows: Section 2 gives the reader an overview of the history of *Gaussian Process* models as well as how they are formulated and how to perform inference with them. Sections 3, 4 describe the GP-AR and GP-ARX models for OSA prediction of *Dst* and how to choose their free parameters for better performance.

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 specified as $y = f(\mathbf{x}) + \epsilon$ where $f(.) : \mathbb{R}^d \rightarrow \mathbb{R}$ is an unknown scalar function of a multidimensional input vector $\mathbf{x} \in \mathbb{R}^d$, d is the dimensionality of the input space, 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 are listed in table 1.

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

2.3. Model Selection

Given a GP model with a kernel function K_θ, the problem of model selection consists of finding appropriate values for the kernel hyper-parameters $\theta = (\theta_1, \theta_2, \cdots, \theta_i)$. In order to assign a value to θ, we must define an objective function which represents our confidence that the GP model built from a particular value of θ is the best performing model. Since GP models encode assumptions about the probability distribution of the output data $\mathbf{y}$ given inputs $\mathbf{X}$, it is natural to use the negative log-likelihood of the training data as a model selection criterion.

$$
\begin{aligned}
Q(\theta) &= -log(p(\mathbf{y}|\mathbf{X}, K_\theta)) \\
&= -\frac{1}{2}\mathbf{y}^\intercal(\mathbf{K}_\theta + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}|\mathbf{K}_\theta + \sigma_n^2\mathbf{I}| - \frac{n}{2}log(2\pi) \\
\mathbf{K}_\theta &= [K_\theta(\mathbf{x}_i, \mathbf{x}_j)]_{n\times n}
\end{aligned}
$$

The model selection problem can now be expressed as the minimization problem shown below.

$$\theta^* = arg\min_\theta \; Q(\theta)$$

The objective function $Q(\theta)$ in the general case can have multiple local minima, and evaluating the value of $Q(.)$ at any given θ requires inversion of the matrix $\mathbf{K}_\theta + \sigma_n^2\mathbf{I}$ which has a time complexity $O(n^3)$ as noted above. In the interest of saving computational cost, one cannot use exhaustive search through the domain of the hyper-parameters to inform our choice for θ. Some of the techniques used for model selection in the context of GPR include.

1. Grid Search: Construct a grid of values for θ as the cartesian product of one dimensional grids for each θ_i, evaluate $Q(.)$ at each such grid point and choose the configuration which yields minimum value of $Q(.)$.

2. Coupled Simulated Annealing: Introduced in *Xavier-De-Souza et al.* [2010], it follows the same procedure as *grid search*, but after evaluation of the objective $Q(.)$ on the grid, each grid point is iteratively mutated in a random walk fashion. This mutation is accepted or rejected according to the new value of $Q(.)$ as well as its value on the other grid points. This procedure is iterated until some stop criterion is reached.

3. Maximum Likelihood: This technique as outlined in *Rasmussen and Williams* [2005] is a form of *gradient descent*. It involves starting with an initial guess for θ and iteratively improving it by calculating the gradient of $Q(.)$ with respect to θ. Although this method seems intuitive, it introduces an extra computational cost of calculating the gradient of $Q(\theta)$ with respect to each θ_i in every iteration and applying this method can sometimes lead to overfitting of the GPR model to the training data.

3. One Step Ahead Prediction

Below in equations (15) - (17) 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 $B_z(t)$.

$$Dst(t) = f(\mathbf{x}_{t-1}) + \epsilon \tag{15}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{16}$$

$$f(x_t) \sim \mathcal{GP}(m(\mathbf{x}_t), K_{osa}(\mathbf{x}_t, \mathbf{x}_s)) \tag{17}$$

$$\tag{18}$$

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 and IMF B_z.

In this model we choose distinct time lags p, p_v and p_b for Dst, V and B_z respectively.

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

3.3. Choice of Mean Function

Mean functions in GPR models encode trends in the data, they are the baseline predictions the model falls back to in case the training and test data have little correlation as predicted by the kernel function. If there is no prior knowledge about the function to be approximated, *Rasmussen and Williams* [2005] state that it is perfectly reasonable to choose $m(\mathbf{x} = 0)$ as the mean function, as long as the target values are normalized. In the case of the *Dst* time series, it is known that the so called *persistence model* $\hat{D}st(t) = Dst(t-1)$ performs quite well in the context of OSA prediction. We therefore choose the *persistence model* as the mean function in our OSA Dst models.

3.4. Choice of Kernel

In this study, we construct Gaussian Process regression models with a combination of the *maximum likelihood perceptron* kernel and *student's T* kernel as shown in equations (19). The *maximum likelihood perceptron* kernel is the *Gaussian Process* equivalent of a single hidden layer feed-forward neural network model as demonstrated in *Neal* [1996].

$$K_{osa}(\mathbf{x}, \mathbf{y}) = K_{mlp}(\mathbf{x}, \mathbf{y}) + K_{st}(\mathbf{x}, \mathbf{y}) \tag{19}$$

$$K_{mlp}(\mathbf{x}, \mathbf{y}) = sin^{-1}\left(\frac{w\mathbf{x}^{\intercal}\mathbf{y} + b}{\sqrt{w\mathbf{x}^{\intercal}\mathbf{x} + b + 1}\sqrt{w\mathbf{y}^{\intercal}\mathbf{y} + b + 1}}\right) \tag{20}$$

$$K_{st}(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + ||\mathbf{x} - \mathbf{y}||_2^d} \tag{21}$$

4. Experiments

Training

We selected OMNI data sections 00:00 January 3 2010 to 23:00 January 23 2010 and 20:00 August 5 2011 to 22:00 August 6 2011 for training the GP-AR and GP-ARX models. The first training data section consists of ambient fluctuations of *Dst* while the second contains a geomagnetic storm.

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 243 which is still very much below the computational limits of the method and in our case solving equation 13 on a laptop computer takes less than a second for the training set considered in our analysis.

Selection

In order to find appropriate values of the hyper-parameters of the chosen kernel K_{osa}, we apply *grid search*, *coupled simulated annealing* and *maximum likelihood* methods. We fix the parameters d and σ^2 of K_{st} and model noise to values 0.01 and 0.2 respectively, the remaining parameters w and b are kept free to be calculated by model selection. Table 2 summarizes the settings used to run each model selection procedure.

Validation

Apart from selecting the kernel parameters, one also needs to choose appropriate values for the auto-regressive orders p in the case of GP-AR and p, p_v, p_b in the case of GP-ARX. For this purpose we use a set of 24 storm events listed in table 4 and for every assignment of values

to the model order, we perform model selection with the routines in table 2 and record the performance on this validation set.

For measuring performance of model instances on the validation set storm events, the following metrics are calculated.

1. The mean absolute error.

$$MAE = \sum_{t=1}^{n} \left|(Dst(t) - \hat{D}st(t))\right| /n \tag{22}$$

2. The root mean square error.

$$RMSE = \sqrt{\sum_{t=1}^{n} (Dst(t) - \hat{D}st(t))^2 /n} \tag{23}$$

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

$$CC = Cov(Dst, \hat{D}st) / \sqrt{Var(Dst)Var(\hat{D}st)} \tag{24}$$

In the case of GP-AR we let the model order p vary from 5 to 12 while for GP-ARX we vary the total model order $p_t = p + p_v + p_b$ vary from 3 to 12 and for each p_t evaluate every possible combination of p, p_v and p_b such that $p_t = p + p_v + p_b$ and $p, p_v, p_b > 0$.

Evaluation

After selecting the best performing GP-AR and GP-ARX models in the validation phase, we test and compare the performance of these models with the predictions generated from the *persistence model* $\hat{D}st(t) = Dst(t-1)$, on a set of 63 storm events occurring between 1998 and 2006 as given in table 5, which is the same list of storm events as used in *Ji et al.* [2012].

5. Results

Figures 1 and 2 show how the mean absolute error and coefficient of correlation as calculated on the validation set storm events of table 4, vary with increasing model order for GP-AR and

GP-ARX. The results are represented as boxplots, in which a rectangle is drawn to represent the second and third quartiles, with a vertical line inside to indicate the median value while outlying points are shown as dots. In both cases, the predictive performance first improves and then stagnates or worsens with increasing model order.

Figures 3 and 4 break down the results for GP-ARX by the model selection routine used. Apart from the general trend observed in 1 and 2, we also observe that *grid search* and *coupled simulated annealing* give superior performance as compared to gradient based *maximum likelihood*.

From the validation results, we can chose the model order which yields the best performance, for GP-AR it is $p_t = 6$ while for GP-ARX it is $p_t = 11$. Further examination of the validation results shows that in the scheme $p_t = 11$ choosing $p = 7, p_v = 1, p_b = 3$ gives superior results.

After choosing the best performing GP-AR and GP-ARX models, we calculate their performance on the test set of table 5. The results of these model evaluations are summarized in table 3, the GP-AR and GP-ARX models improve upon the performance of the *persistence model*.

Figures 5, 6 and 7 show OSA predictions of the GP-ARX model with $\pm\sigma$ error bars for three storm events in the time period between 1998 and 2003. The GP-ARX model gives accurate predictions along with plausible error bars around its mean predictions.

6. Conclusions

In this paper, we describe a flexible and expressive methodology for generating probabilistic forecasts of the *Dst* index. We proposed two *Gaussian Process* auto-regressive models, *GP-ARX* and *GP-AR*, to generate hourly predictions and their associated error bars. We also describe how to carry out model selection and validation of GP-AR and GP-ARX models.

Our results can be summarized as follows.

1. *Persistence* model plays an important role in the model building and 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.

2. *Gaussian Process* AR and ARX models give encouraging benefits in OSA prediction. 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 243 instances, while the validation set contains 782 instances.

3. 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.

4. The *Gaussian Process* regression framework described in this study can also be extended to multiple hour ahead prediction of *Dst*, which is currently a work in progress.

Acknowledgments. We acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb (or CDAWeb or ftp) service, and OMNI data. Simon Wing acknowledges supports from CWI and NSF Grant AGS-1058456 and NASA Grants (NNX13AE12G, NNX15AJ01G, NNX16AC39G).

References

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

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

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

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

Berlinet, A., and C. Thomas-Agnan (2004), *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, 355 pp., Springer US, doi:10.1007/978-1-4419-9096-9.

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

Billings, S. A., S. Chen, and M. J. Korenberg (1989), Identification of mimo non-linear systems using a forward-regression orthogonal estimator, *International Journal of Control*, *49*(6), 2157–2189, doi:10.1080/00207178908559767.

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

Boynton, R. J., M. A. Balikhin, S. A. Billings, A. S. Sharma, and O. A. Amariutei (2011a), Data derived narmax dst model, *Annales Geophysicae*, *29*(6), 965–971, doi:10.5194/angeo-29-965-2011.

Boynton, R. J., M. A. Balikhin, S. A. Billings, H. L. Wei, and N. Ganushkina (2011b), 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, doi:10.1029/2010JA015505.

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 (2013), The analysis of electron fluxes at geosynchronous orbit employing a narmax approach, *Journal of Geophysical Research: Space Physics*, *118*(4), 1500–1513, doi:10.1002/jgra.50192.

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

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

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

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

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

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

Kissinger, J., R. L. McPherron, T.-S. Hsu, and V. Angelopoulos (2011), Steady magnetospheric convection and stream interfaces: Relationship over a solar cycle, *Journal of Geophysical Research: Space Physics*, *116*(A5), n/a–n/a, doi:10.1029/2010JA015763, a00I19.

Krige, d. g. (1951), *A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand*, publisher not identified.

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

McPherron, R. L., T. Terasawa, and A. Nishida (1986), Solar wind triggering of substorm expansion onset, *Journal of geomagnetism and geoelectricity*, *38*(11), 1089–1108, doi: 10.5636/jgg.38.1089.

McPherron, R. L., G. Siscoe, N. U. Crooker, and N. Arge (2013), *Probabilistic Forecasting of the Dst Index*, pp. 203–210, American Geophysical Union, doi:10.1029/155GM22.

Neal, R. M. (1996), *Bayesian Learning for Neural Networks*, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Newell, P., K. Liou, J. Gjerloev, T. Sotirelis, S. Wing, and E. Mitchell (2016), Substorm probabilities are best predicted from solar wind speed, *Journal of Atmospheric and Solar-Terrestrial Physics*, *146*, 28 – 37, doi:http://dx.doi.org/10.1016/j.jastp.2016.04.019.

O'Brien, T. P., and R. L. McPherron (2000), 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, doi:10.1029/1998JA000437.

Pallocchia, G., E. Amata, G. Consolini, M. F. Marcucci, and I. Bertello (2006), Geomagnetic Dst index forecast based on IMF data only, *Annales Geophysicae*, *24*(3), 989–999.

Rasmussen, C. E., and C. K. I. Williams (2005), *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press.

Scholkopf, B., and A. J. Smola (2001), *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA.

Tao, T. (2011), *An Introduction to Measure Theory*, Graduate studies in mathematics, American Mathematical Society.

Wang, C. B., J. K. Chao, and C.-H. Lin (2003), 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, doi:10.1029/2003JA009851, 1341.

Wang, Y.-M., and N. R. Sheeley, Jr. (1990), Solar wind speed and coronal flux-tube expansion, *Astrophys. J.,* , *355*, 726–732, doi:10.1086/168805.

Wing, S., J. R. Johnson, J. Jen, C.-I. Meng, D. G. Sibeck, K. Bechtold, J. Freeman, K. Costello, M. Balikhin, and K. Takahashi (2005), Kp forecast models, *Journal of Geophysical Research: Space Physics*, *110*(A4), n/a–n/a, doi:10.1029/2004JA010500, a04203.

Xavier-De-Souza, S., J. A. K. Suykens, J. Vandewalle, and D. Bolle (2010), Coupled simulated annealing, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *40*(2), 320–335, doi:10.1109/TSMCB.2009.2020435.

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

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

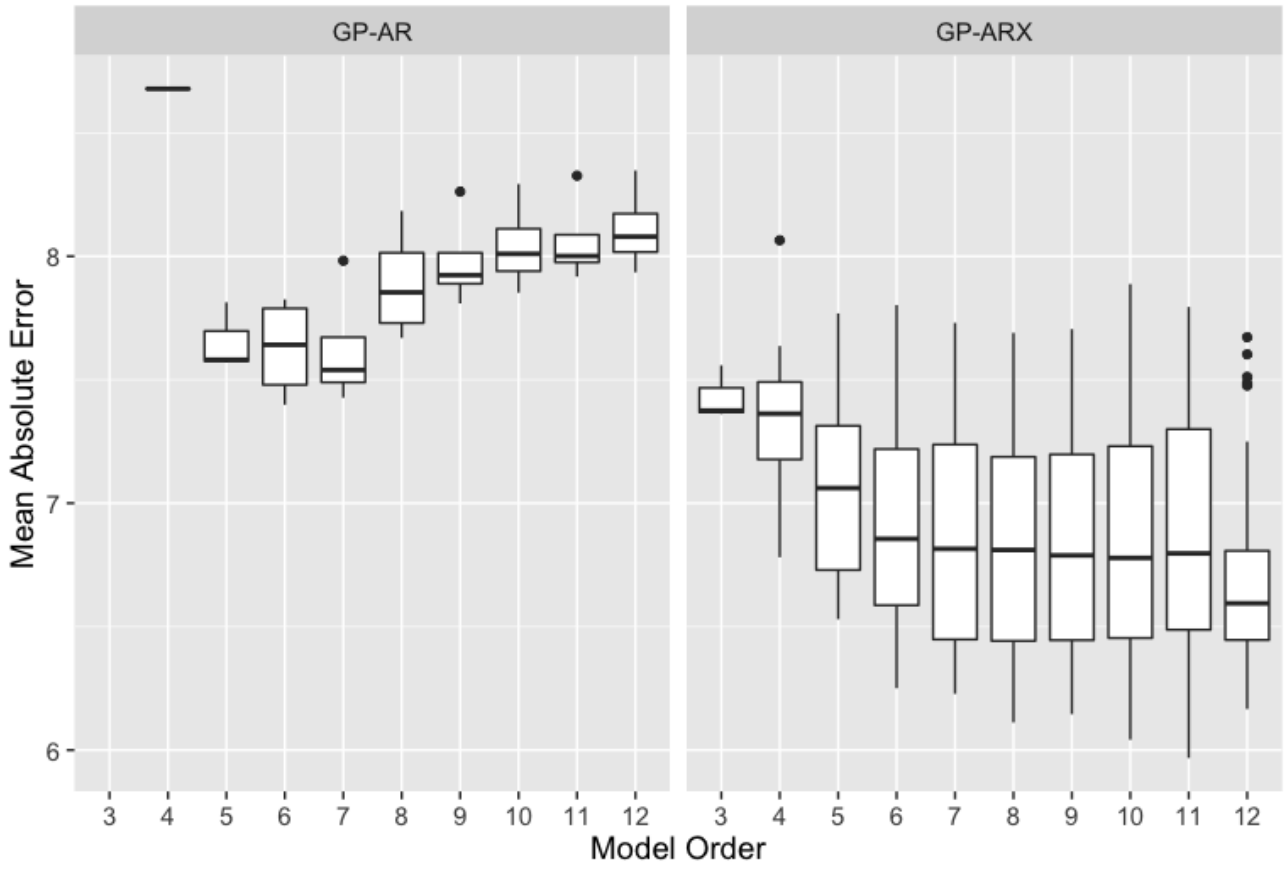

Figure 1. Mean Absolute Error on validation set storms vs model order for GP-AR and GP-ARX.

Key: Rectangle borders represent the second and third quartiles, with a horizontal line inside to indicate the median value while outlying points are shown as dots

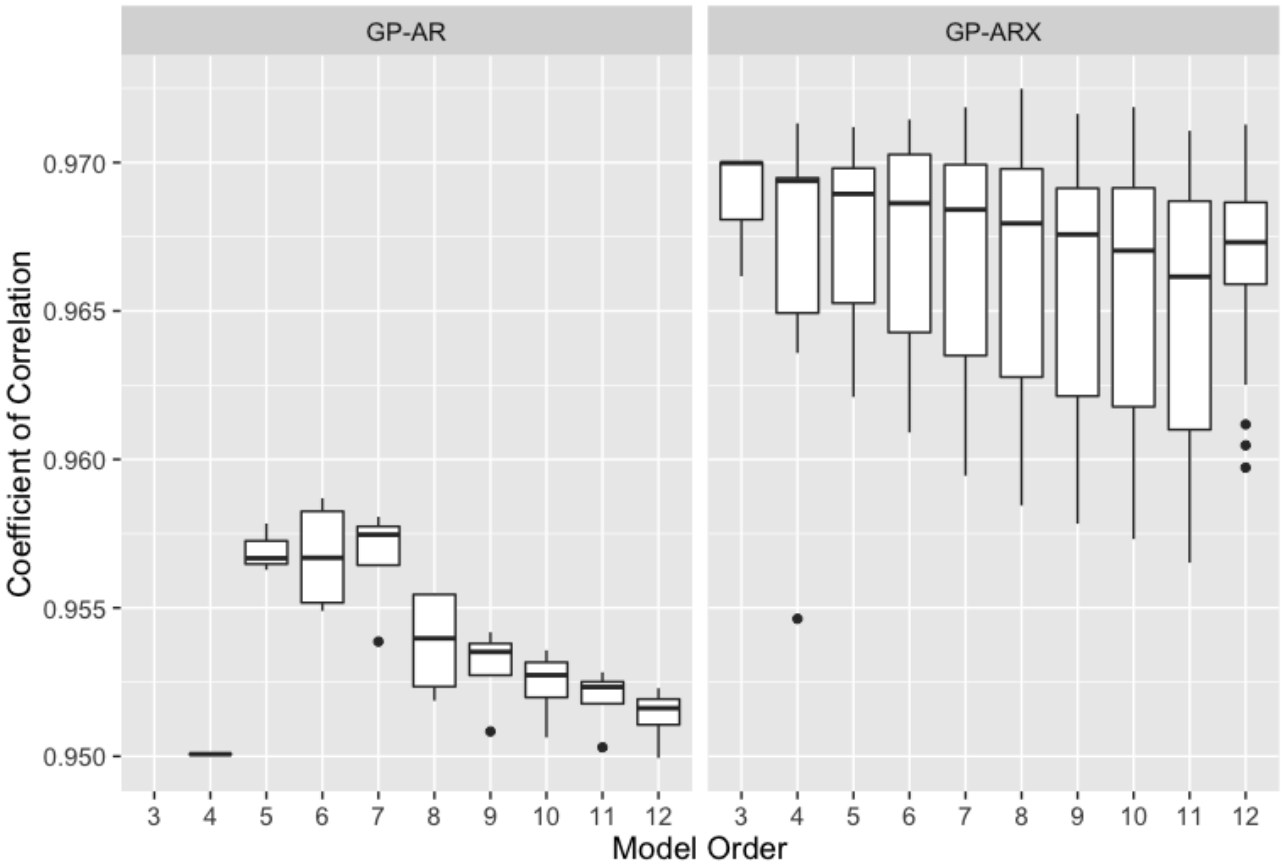

Figure 2. Coefficient of Correlation on validation set storms vs model order for GP-AR and GP-ARX

Key: Rectangle borders represent the second and third quartiles, with a horizontal line inside to indicate the median value while outlying points are shown as dots

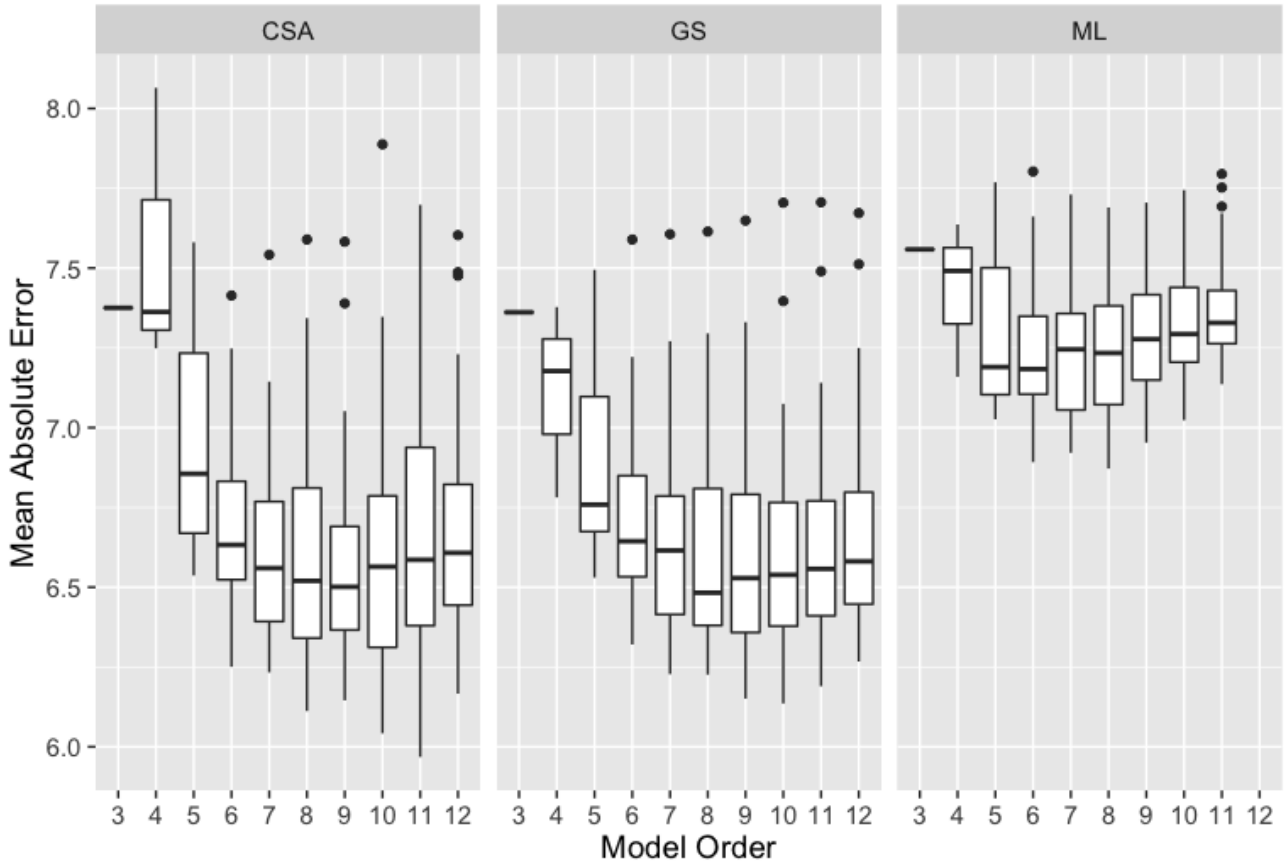

Figure 3. Mean Absolute Error on validation set storms vs model order for GP-AR and GP-ARX for *CSA*, *GS* and *ML* model selection routines

Rectangle borders represent the second and third quartiles, with a horizontal line inside to indicate the median value while outlying points are shown as dots

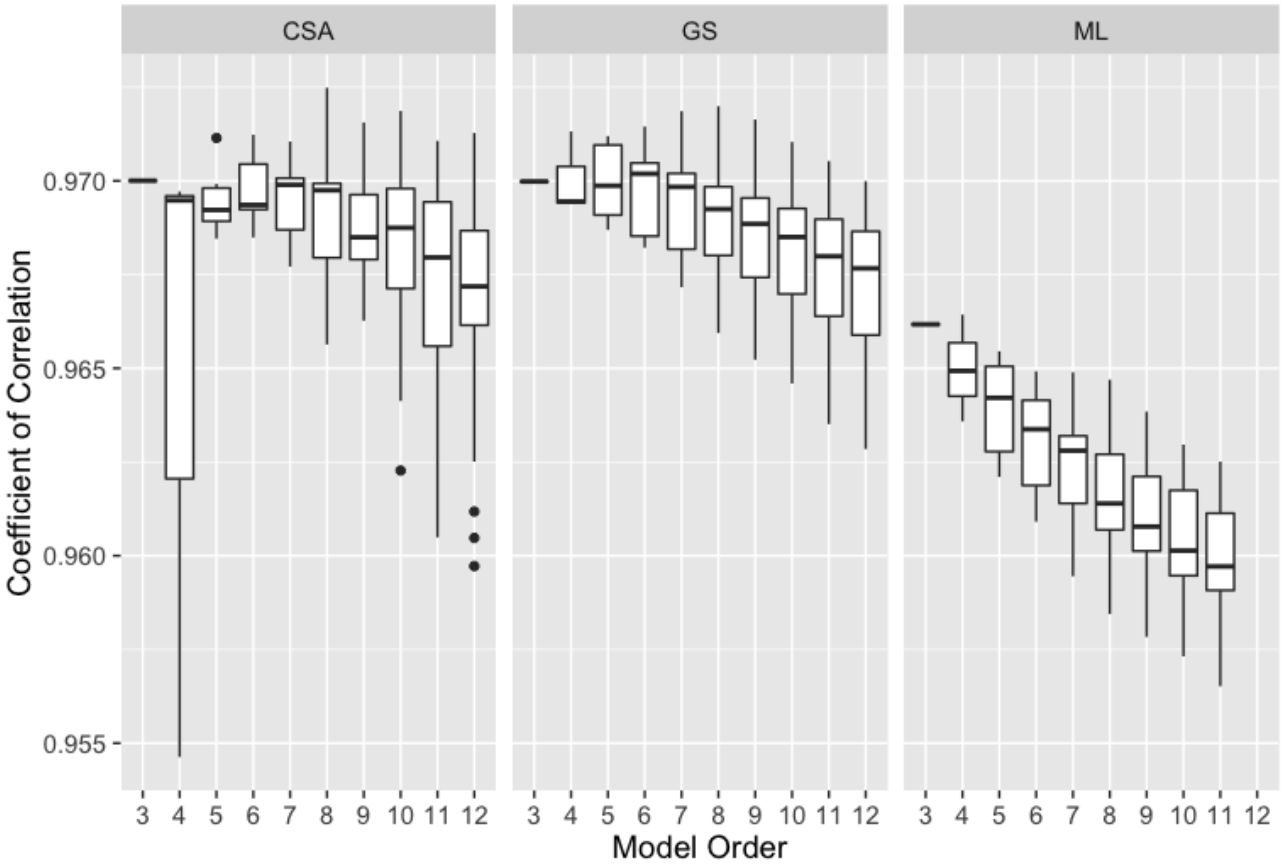

Figure 4. Coefficient of Correlation on validation set storms vs model order for GP-AR and GP-ARX for *CSA*, *GS* and *ML* model selection routines

Rectangle borders represent the second and third quartiles, with a horizontal line inside to indicate the median value while outlying points are shown as dots

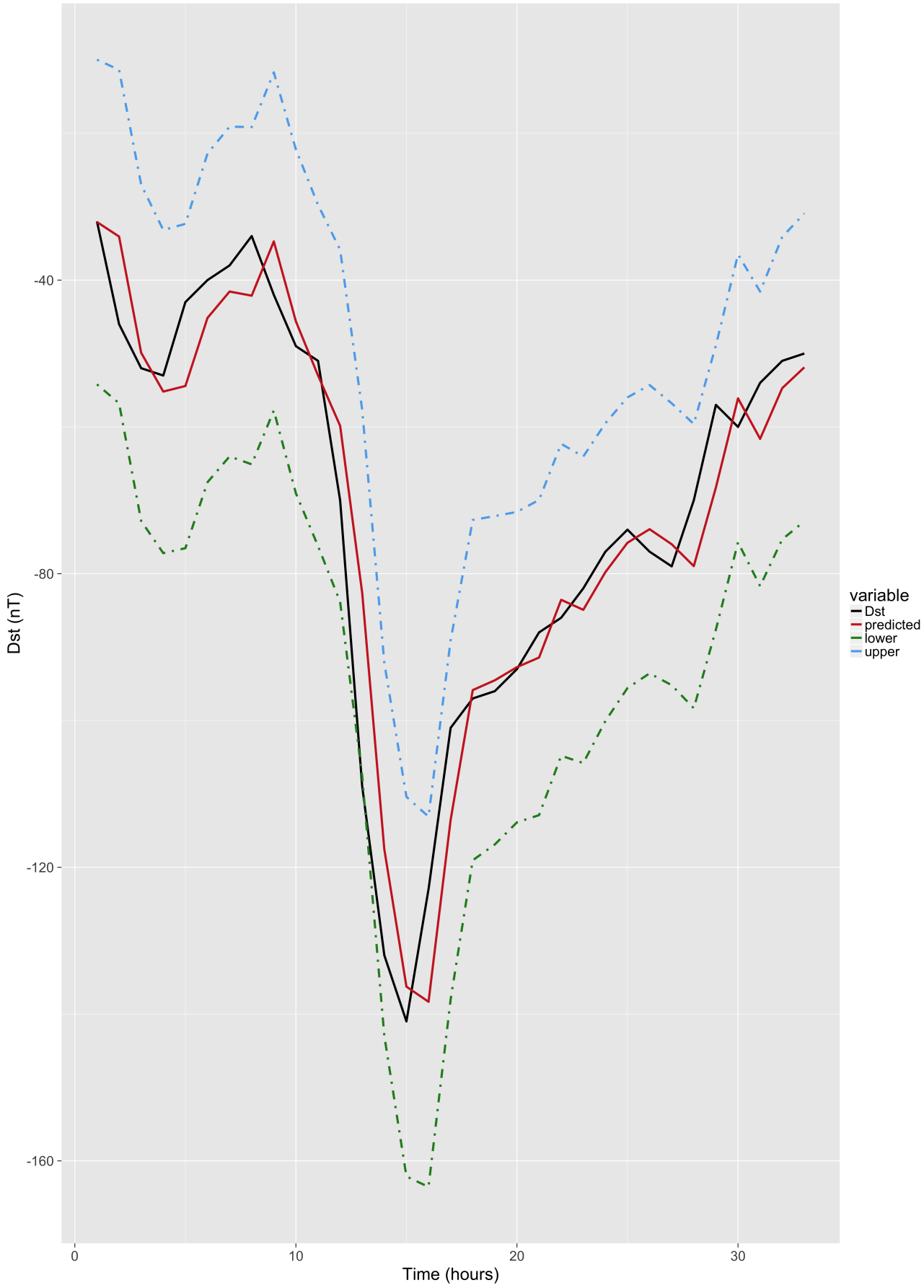

Figure 5. OSA Predictions with $\pm\sigma$ error bars for event: 2003/06/17 19:00 to 2003/06/19 03:00

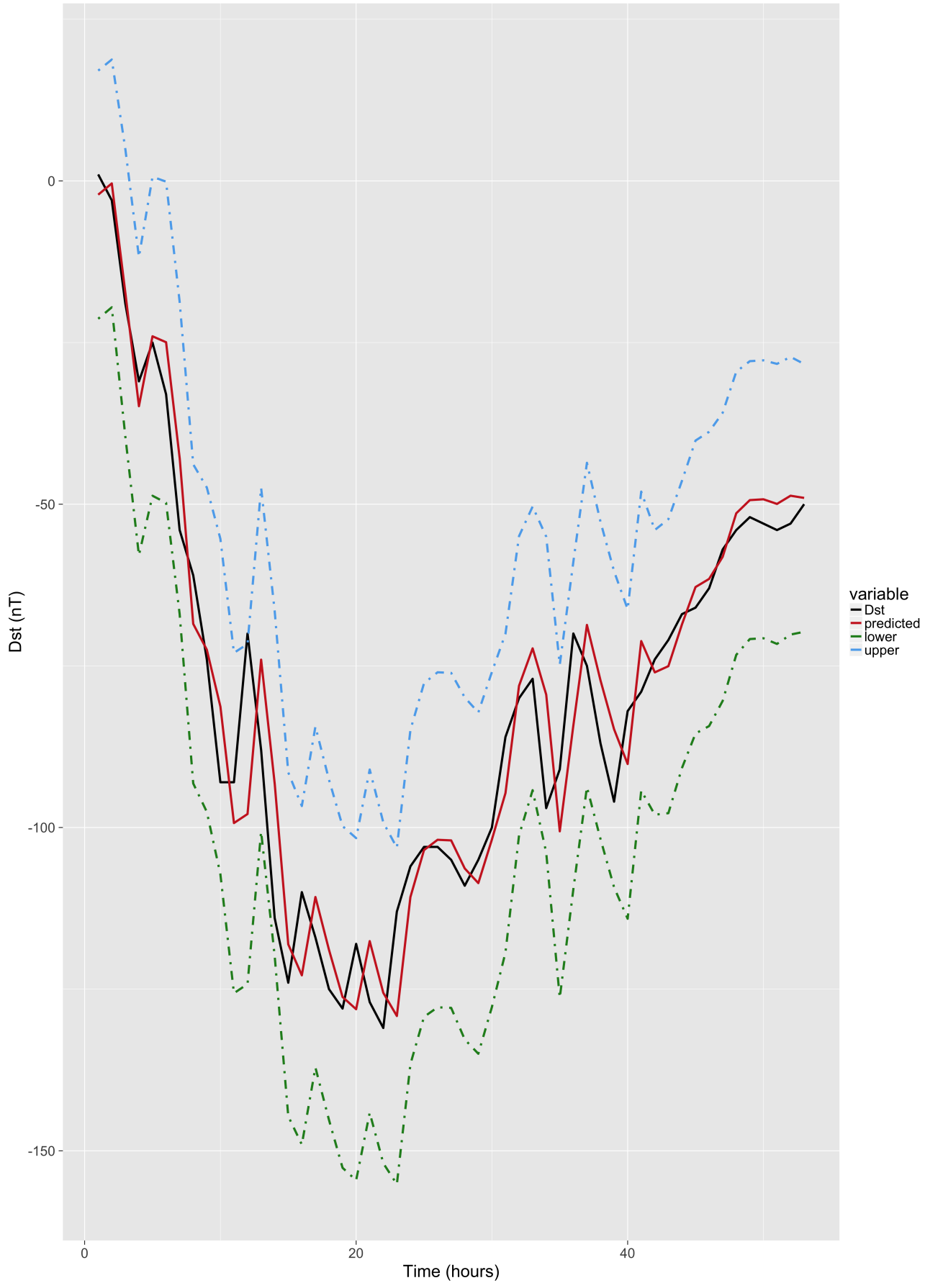

Figure 6. OSA Predictions with $\pm\sigma$ error bars for event: 1998/11/13 00:00 to 1998/11/15 04:00

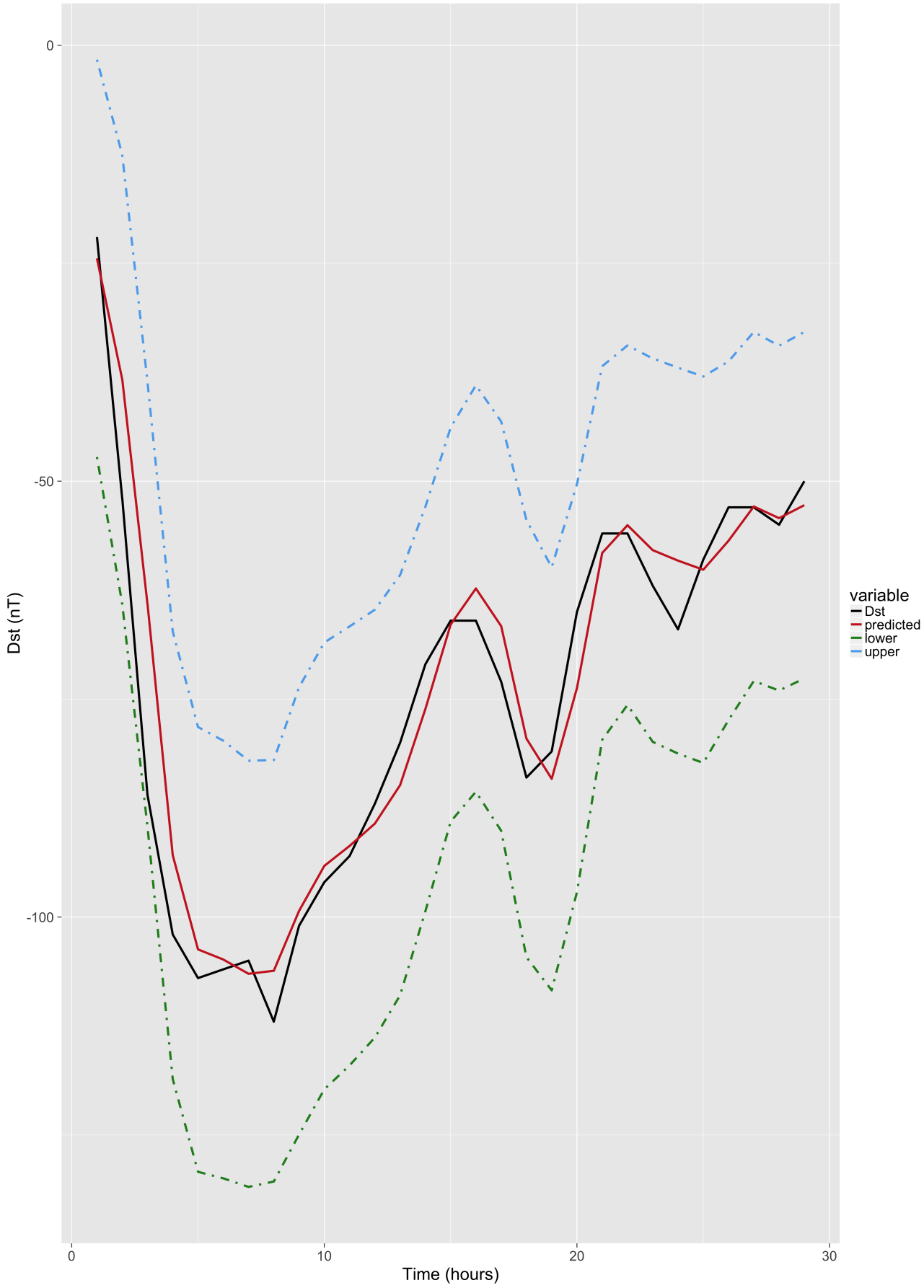

Figure 7. OSA Predictions with $\pm\sigma$ error bars for event: 1999/01/13 16:00 to 1999/01/14 20:00

Table 1. Popular Kernel functions used in GPR models

Name	Expression	Hyperparameters
Radial Basis Function (RBF)	$\frac{1}{2}exp(-\|\|\mathbf{x}-\mathbf{y}\|\|^2/l^2)$	$l \in \mathbb{R}$
Polynomial	$(\mathbf{x}^\intercal\mathbf{y}+b)^d$	$b \in \mathbb{R}, d \in \mathbb{N}$
Laplacian	$exp(-\|\|\mathbf{x}-\mathbf{y}\|\|_1/\theta)$	$\theta \in \mathbb{R}^+$
Student's T	$1/(1+\|\|\mathbf{x}-\mathbf{y}\|\|_2^d)$	$d \in \mathbb{R}^+$
Maximum Likelihood Perceptron	$sin^{-1}(\frac{w\mathbf{x}^\intercal\mathbf{y}+b}{\sqrt{w\mathbf{x}^\intercal\mathbf{x}+b+1}\sqrt{w\mathbf{y}^\intercal\mathbf{y}+b+1}})$	$w, b \in \mathbb{R}^+$

Table 2. Settings of model selection procedures

Procedure	Grid Size	Step	Max Iterations
Grid Search	10	0.2	NA
Coupled Simulated Annealing	4	0.2	30
Maximum likelihood	NA	0.2	150

Table 3. Evaluation results for models on storm events listed in table 5

Model	Mean Absolute Error	Root Mean Square Error	Coefficient of Correlation
GP-ARX	7.252	11.93	0.972
GP-AR	8.37	14.04	0.963
Persistence	9.182	14.94	0.957

Table 4. Storm events used for model selection of GP-AR and GP-ARX

Event Id	Start Date	Start Hour	End Date	End Hour	min. Dst
1	1995/03/26	0500	1995/03/26	2300	107
2	1995/04/07	1300	1995/04/09	0900	149
3	1995/09/27	0100	1995/09/28	0400	108
4	1995/10/18	1300	1995/10/19	1400	127
5	1996/10/22	2200	1996/10/23	1100	105
6	1997/04/21	1000	1997/04/22	0900	107
7	1997/05/15	0300	1997/05/16	0000	115
8	1997/10/10	1800	1997/10/11	1900	130
9	1997/11/07	0000	1997/11/07	1800	110
10	1997/11/22	2100	1997/11/24	0400	108
11	2005/06/12	1700	2005/06/13	1900	106
12	2005/08/31	1200	2005/09/01	1200	122
13	2006/12/14	2100	2006/12/16	0300	162
14	2011/09/26	1400	2011/09/27	1200	101
15	2011/10/24	2000	2011/10/25	1400	132
16	2012/03/08	1200	2012/03/10	1600	131
17	2012/04/23	1100	2012/04/24	1300	108
18	2012/07/15	0100	2012/07/16	2300	127
19	2012/09/30	1300	2012/10/01	1800	119
20	2012/10/08	0200	2012/10/09	1700	105
21	2012/11/13	1800	2012/11/14	1800	108
22	2013/03/17	0700	2013/03/18	1000	132
23	2013/05/31	1800	2013/06/01	2000	119
24	2014/02/18	1500	2014/02/19	1600	112

Table 5. Storm events used to evaluate GP-AR and GP-ARX models

Event Id	Start Date	Start Time	End Date	End Time	min. Dst
1	1998/02/17	1200	1998/02/18	1000	-100
2	1998/03/10	1100	1998/03/11	1800	-116
3	1998/05/04	0200	1998/05/05	0200	-205
4	1998/08/26	1000	1998/08/29	0700	-155
5	1998/09/25	0100	1998/09/26	0000	-207
6	1998/10/19	0500	1998/10/20	0800	-112
7	1998/11/09	0300	1998/11/10	1600	-142
8	1998/11/13	0000	1998/11/15	0400	-131
9	1999/01/13	1600	1999/01/14	2000	-112
10	1999/02/18	0300	1999/02/19	2100	-123
11	1999/09/22	2000	1999/09/23	2300	-173
12	1999/10/22	0000	1999/10/23	1400	-237
13	2000/02/12	0500	2000/02/13	1500	-133
14	2000/04/06	1700	2000/04/08	0900	-288
15	2000/05/24	0100	2000/05/25	2000	-147
16	2000/08/10	2000	2000/08/11	1800	-106
17	2000/08/12	0200	2000/08/13	1700	-235
18	2000/10/13	0200	2000/10/14	2300	-107
19	2000/10/28	2000	2000/10/29	2000	-127
20	2000/11/06	1300	2000/11/07	1800	-159
21	2000/11/28	1800	2000/11/29	2300	-119
22	2001/03/19	1500	2001/03/21	2300	-149
23	2001/03/31	0400	2001/04/01	2100	-387
24	2001/04/11	1600	2001/04/13	0700	-271
25	2001/04/18	0100	2001/04/18	1300	-114
26	2001/04/22	0200	2001/04/23	1500	-102
27	2001/08/17	1600	2001/08/18	1600	-105
28	2001/09/30	2300	2001/10/02	0000	-148
29	2001/10/21	1700	2001/10/24	1100	-187
30	2001/10/28	0300	2001/10/29	2200	-157
31	2002/03/23	1400	2002/03/25	0500	-100
32	2002/04/17	1100	2002/04/19	0200	-127
33	2002/04/19	0900	2002/04/21	0600	-149
34	2002/05/11	1000	2002/05/12	1600	-110
35	2002/05/23	1200	2002/05/24	2300	-109
36	2002/08/01	2300	2002/08/02	0900	-102
37	2002/09/04	0100	2002/09/05	0000	-109
38	2002/09/07	1400	2002/09/08	2000	-181
39	2002/10/01	0600	2002/10/03	0800	-176
40	2002/10/03	1000	2002/10/04	1800	-146
41	2002/11/20	1600	2002/11/22	0600	-128
42	2003/05/29	2000	2003/05/30	1000	-144
43	2003/06/17	1900	2003/06/19	0300	-141
44	2003/07/11	1500	2003/07/12	1600	-105
45	2003/08/17	1800	2003/08/19	1100	-148
46	2003/11/20	1200	2003/11/22	0000	-422
47	2004/01/22	0300	2004/01/24	0000	-149
48	2004/02/11	1000	2004/02/12	0000	-105
49	2004/04/03	1400	2004/04/04	0800	-112
50	2004/07/22	2000	2004/07/23	2000	-101
51	2004/07/24	2100	2004/07/26	1700	-148
52	2004/07/26	2200	2004/07/30	0500	-197
53	2004/08/30	0500	2004/08/31	2100	-126
54	2004/11/07	2100	2004/11/08	2100	-373
55	2004/11/09	1100	2004/11/11	0900	-289
56	2004/11/11	2200	2004/11/13	1300	-109
57	2005/01/21	1800	2005/01/23	0500	-105
58	2005/05/07	2000	2005/05/09	1000	-127
59	2005/05/29	2200	2005/05/31	0800	-138
60	2005/06/12	1700	2005/06/13	1900	-106
61	2005/08/31	1200	2005/09/01	1200	-131
62	2006/04/13	2000	2006/04/14	2300	-111
63	2006/12/14	2100	2006/12/16	0300	-147