PRediction Of Geospace Radiation Environment and Solar wind parameterS

Work Package 3
Forecast of the evolution of geomagnetic indices

Deliverable 3.7
GMN and bi-linear *Dst* and *Kp* models: Development, testing and comparison of model outputs

P. Wintoft, M. Wik, J. Katkalov
IRF,
R. Boynton, H.-L. Wei, S. Walker,
M. Balikhin, R. Erdelyi
USFD,
V. Yatsenko, O. Cheremnykh, O. Semeniv,
J. Krivickaya, S. Ivanov, I. Mulko, A. Bespalova
SRI

July 13, 2018

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 637302.

Document Change Record

Issue	Date	Author	Details
v1	June 26, 2018	P. Wintoft	First draft.
v2	June 29, 2018	S. Walker	updates to first draft, addition of model descriptions.
v3	July 2, 2018	S. Walker	restructure, added model polynomials
v4	July 10, 2018	S. Walker	Incorporate comments from HLW
v5	July 12, 2018	S. Walker	Incorporate comments from OC

Contents

Summary

This deliverable is an extension of the PROGRESS project according to Amendment 23. It encapsulates the work originally proposed by the participant SRI as part of Tasks T3.4, T3.5, and T3.6 of Work Package 3. During the main period (2015-01-01 to 2017-12-31) of the project, SRI began to develop their forecast models for the geomagnetic indices Kp, Dst, and AE. Unfortunately, these models were never delivered. During the Project extension phase (2018-01-01 to 2018-07-31) the PROGRESS participants were given extra time to complete the development of these models and deliver the resulting products.

SRI originally promised to develop models for geomagnetic indices based on the following methodologies:

- a recursive, robust bilinear dynamical model
- a Guaranteed NARMAX Model

The Commission proposed that the Project reallocate some of the tasks allotted to SRI to other participants with the knowledge and skills to complete them. Following discussions with SRI and IRF it was agreed that

- SRI will continue to develop their models for Kp, Dst, and AE based on their Guaranteed NARMAX Model,
- USFD will develop models of the Kp and Dst indices based on the recursive, robust bilinear dynamical methodology.
- USFD will study the Lyapunov Exponents of the Dst data set in order to determine the forecast horizon
- IRF will investigate the performance of the models, perform an inter-comparison

These subtasks were incorporated into a new task, T3.7, within WP 3.

This document is a new deliverable, D3.7, which reports on the modelling methodologies employed, the resulting forecasts, and compares the solar wind driven prediction models of the Kp, Dst, and AE indices developed by IRF, USFD, and SRI.

Acronyms

ACE	Advanced Composition Explorer
ANOVA	ANalysis Of VAriance
ANN	Artificial Neural Network
AR	AutoRegression
ARX	AutoRegression with eXogeneous inputs
ASCII	American Standard Code for Information Interchange
DSCOVR	Deep Space
BS	Brier scoreClimate Observatory
DoA	Description of Action
GE	Guaranteed Estimation
GFZ	GeoForschungsZentrum
GNM	Guaranteed NARMAX Model
GP	Genetic Programming
GSFC	Goddard Space Flight Center
IMF	Interplanetary magnetic field
IRF	Institutet för rymdfysik (Swedish Institute of Space Physics)
L1	Lagrange point 1 (between Earth and Sun)
LC	Linear correlation coefficient
LLE	Largest Lyapunov exponent
MPO	Model predicted
NARMAX	Nonlinear AutoRegression Moving Average with eXogeneous inputs
NARX	Nonlinear AutoRegression with eXogeneous inputs
NASA	National Aeronautic and Space Administration
NCEI	National Centers for Environmental Information
NOAA	National Oceanographic and Atmospheric Administration
NRT	Near-real time
NSSDC	National Space Science Data Center
OSA	One step ahead
PROGRESS	PRediction Of Geospace Radiation Environment and Solar wind parameterS
RMSE	Root mean square error
SQL	Structured Query Language
SRI	Space Research Institute of National Academy of Sciences of Ukraine and State Space Agency of Ukraine, Ukraine
RRBDM	Recursive Robust Bilinear Dynamical Model
SPDF	Space Physics Data Facility
SS	Skill score
SW	Solar wind
SWPC	Space Weather Prediction Center
USFD	University of Sheffield
URL	Uniform Resource Locator
UT	Universal time
WDC	World Data Center

1 Introduction

Several different prediction models that build on different modelling approaches have been developed in the PROGRESS project. All models rely on algorithms that derive functions from observed data that maps from solar wind data to geomagnetic indices. The models developed build on neural networks (NN) (IRF), non-linear autoregressive moving-average with exogenous inputs (NARMAX) (USFD), regression modelling (RM) to construct NARX models (SRI), and genetic programming (GP) (SRI).

This document describes

- the methods involved in the development of these models,
- examples of their forecasts,
- a comparison of the performances of the models developed within PROGRESS using various statistical measures.

2 Data sources

For the solar wind there exists several different data sources that can be used depending on the requirements. As this report is concerned with testing and comparing the model outputs we do not consider real-time data.

2.1 OMNI

There exists two different OMNI sets, a high resolution set with 1-minute data and a low-resolution set with 1-hour data. The low resolution set has been used for the model developments. The solar wind data in the OMNI set come from many different spacecraft, but from 1998 and onwards it is dominated by data from the ACE spacecraft. The solar wind data have been propagated from spacecraft location to a point just upstream of the Earth.

2.2 ACE L2

The ACE Level 2 (L2) data are verified science level data. Different data products exist but for the models here the 64 second merged plasma and magnetic field data have been used. The ACE spacecraft moves in an orbit around the Lagrange 1 (L1) point approximately 1.5 million km upstream from Earth.

2.3 Kp index

GFZ provides both nowcast (real-time) and definitive Kp where the definitive data typically are available up to the past month from https://www.gfz-potsdam.de/en/kp-index/. For this study the definitive data are used.

2.4 *Dst* index

WDC-Kyoto provides the *Dst* index at three different processing levels: final before 2015 (http://wdc.kugi.kyoto-u.ac.jp/dst_final/index.html), provisional 2015–2016 (http://wdc.kugi.kyoto-u.ac.jp/dst_provisional/index.html), and real-time from 2017 (http://wdc.kugi.kyoto-u.ac.jp/dst_realtime/i as of May 2018. Thus, about 5 months of data used come from the provisional set while the rest are real-time data. The real-time data is a live data set and past values will change with time due to updates of reported data from the four observatories (Kakioka, Honolulu, San Juan, and Hermanus) and due updated quiet level estimates. The updated real-time *Dst* is the best at the time of derivation (priv. comm. M. Nosé).

2.5 *AE* index

WDC-Kyoto also provides the *AE* index as either a quick-look (quasi-real time) product for monitoring, diagnostic, and forecasting purposes only, provisional values , or final data products.

Quick-look *AE* products are released with the following provisos.

- Values are derived from raw unverified data and often contain spikes.
- Values may be revised as more data become available or base lines are corrected.
- Values will be replaced by provisional and then final values at a later date.

They are available for the period from January 2018 at http://wdc.kugi.kyoto-u.ac.jp/ae_realtime/index.htm

Provisional *AE* values are available for the period January 1996 to March 2018 at http://wdc.kugi.kyoto-u.ac.jp/ae_provisional/index.html.

Final *AE* values are generally available for the period 1975 to 1988 but there are data gaps at http://wdc.kugi.kyoto-u.ac.jp/aeasy/index.html

3 Models

Here we give a short overview of the models. The name of each model is identified by the acronym of the institute responsible for the model development and the target output. As the model terms have been determined from data it is important to specify what data and which periods have been used for model development so that tests can be performed on independent datasets, which is summarised in Table 1.

This section provides a detailed background to the modelling methodologies used for the development of forecast models for the geomagnetic indices *Kp*, *Dst*, and *AE*. In the framework of the PROGRESS project the methods used are based on Artificial Neural Networks (ANN) and variants of the NARMAX (Nonlinear AutoRegression Moving Average with eXogeneous inputs) systems identification technique.

Table 1: Models with data sources and data periods used for model development.

Model	**Horizon**	**Source**	**Data period**
IRF-Kp-2017	3h	ACE L2	1998 – 2015 except 2001 and 2011
USFD-Kp	3h	OMNI	1998 – 2017
SRI-Kp-GP	3h	OMNI	2006
SRI-Kp-RM	3h	OMNI	1976 – 2008
IRF-Dst-2017	1h	OMNI	1963 – 2015 except 1981, 1996, 2001, and 2008
USFD-Dst	1h	OMNI	2001 – 2002
USFD-Dst	3h	OMNI	2001 – 2002
SRI-Dst-GP	1h	OMNI	2006
SRI-Dst-RM	1h	OMNI	1976 – 2008
SRI-Dst-RM	3h	OMNI	1976 – 2008
IRF-AE-2017		ACE L2	1998 – 2015 except 2001, 2005, and 2013
SRI-AE		OMNI	2013-03-12 – 2013-06-03

3.1 Artificial Neural Networks

The development of methodologies based on Artificial Neural Networks for the forecast of geomagnetic indices has been described in detail in deliverables D3.4 *Kp* and *Dst* models and deliverable D3.5 - *AE* models.

3.2 NARMAX

NARMAX (Nonlinear AutoRegressive Moving Average models with eXogenous inputs) type models (Leontaritis & Billings 1985*a*,*b*, Billings 2013) can capture the dynamics of a nonlinear system, providing both the ability not only to forecast the evolution of a system but also to provide insight into the physical processes underlying the dynamic of the system. A NARMAX model describes the current output of a system as a function of the time lagged input parameter set and previous system output values. This may be expressed in the form (1)

$$y(k) = F[y(k-1), ...y(k-n), u(k-1), ...u(k-n), e(k-1), ...e(k-n)] \qquad (1)$$

where k represents the current measurement time, n the number of time lags, y is the set of output parameters at lags $(k-1)$ to $(k-n)$, u the set of system inputs, and e a set of error terms. $F[]$ represents a nonlinear function, typically either a polynomial, B-spline, or radial basis function. For the purposes of the models developed within PROGRESS $F[]$ is a polynomial function.

The first task within the generalised NARMAX methodology is to determine the structure of the model, i.e. to determine the most significant model parameters. Within the PROGRESS project, there are three different methodologies employed to do this task. An overview of thee methods is given in the next three sections.

3.2.1 Guaranteed NARMAX Method

The Guaranteed NARMAX Model (GNM) views the Sun-Earth system as a nonlinear dynamical system of a black box type with solar wind parameters as the input, and geomagnetic indices as the output. This approach seeks to describe its dynamics and predict its future state using as little a priori knowledge as possible. This is justified by the fact that the current understanding of solar-terrestrial physics is patchy at best. GNM combines the ideas of NARMAX (Leontaritis & Billings 1985*a*) and Guaranteed Estimation (Schweppe 1968) approaches resulting in a variant on the traditional NARMAX model .

The cornerstone idea of the GNM model comes from the Guaranteed Estimation (GE) approach which predicts an interval that contains the true value of the predictand with a guarantee, rather than a single value with or without an error. The justification for such a treatment is that supplying a prediction as a single value with an error implies a stochastic nature of possible deviations from the true value while systems with significant complexity and strong nonlinearity, such as the Sun-Earth system, do not necessarily behave stochastically. The GE approach deals instead with a bounded uncertainty of arbitrary nature. To bind the uncertainty, the training sample is used to build as many constraints as possible, which are then combined to produce a polyhedron in parameter space with a (hopefully) finite volume. This procedure requires the training sample to be representative of the general population, which is possible only to a certain extent. Thus, in practice, "with a guarantee" implies "at a sufficiently high confidence level", which is defined by the Lebesgue measure of the training sample in the state space of the dynamical system considered.

The baseline for this interval is determined using a NARMAX model. Various approaches are possible (Billings 2013). In this particular study polynomial NARX models were identified using algorithms based on the genetic progratmming method of Semeniv (2015) and the regression modelling method of Parnowski (2011). The main difference between these two approaches is in the way they address the structural identification problem. The genetic programming approach views it as an optimization problem, using an algorithm to search for an optimal model structure in a limited space of possible structures. The regression modelling approach follows a more traditional approach and tries to compare all possible models, gradually increasing the complexity of the model.

3.2.2 Genetic programming approach

The Genetic Programming (GP) is a widely used population based iterative optimization technique developed in late 1960s – early 1970s (see e.g. Bosworth et al. 1972). It is an evolutionary computation technique based on the so-called "tree representation" (Koza 1992, Koza et al. 2003). The problem is transformed to the GP by performing certain well-defined steps: the set of terminals for each branch of the model; the set of primitive functions for each branch of the model; the fitness measure; certain parameters for controlling the run; the stop criterion. GP typically starts with a population of randomly generated models (regressors) composed from the available input variables (Koza 1992, Koza et al. 2003, Madar et al. 2005). GP iteratively transforms a population of individuals into a new generation of the population by applying analogues of naturally occurring genetic operations. These operations are applied to individuals selected from the popula-

tion. The individuals are probabilistically selected to participate in the genetic operations based on their fitness value. The iterative transformation of the population is executed in the main generational loop of GP. A population member in GP is a hierarchically structured tree consisting of functions and terminals. The functions and terminals are selected from a set of functions and a set of terminals. For example, the set of operators can contain the basic arithmetic operations: $\{+, -, \times, \div, \sqrt{}\}$. Potential solution may be depicted as a rooted, labelled tree with ordered branches, using operations from the function set and arguments from the terminal set.

Population members represent linear and nonlinear base functions. The parameters are joined to the model after extracting these functions from the tree, and they are determined using the least squares method. One can extract the function terms by decomposing the tree starting from the root. If the set of operators is defined as $F = \{+, \times\}$ and there is a syntactic rule that exchanges the underlying internal nodes to $\times$-type and $+$-type nodes, the algorithm will generate only polynomial models. The most widely used selection strategy is the roulette-wheel selection. In the roulette-wheel selection, every individual (model structure) has a probability to be selected as parent, and this probability is proportional to fitness value ϕ_j, which is a mean of fitness value ϕ_j of the individual j, which can be defined in a number of ways. When an individual is selected for reproduction, three operations can be applied: direct reproduction, mutation and crossover (recombination). The probability of mutation is pm, the probability of crossover is pc, and the probability of direct reproduction is $1 - pm - pc$.

According to the values pm, pc, and ϕ_j (which are set by the investigator) the procedure of model reconstruction is conducted. The direct reproduction puts the selected individual into the new generation without any change. In mutation a random change is performed on the selected tree structure (model) by a random substitution. If an internal element (an operator) is changed to a leaf element (an argument), the structure of tree will change too. In crossover two individuals are selected, and their tree structures are divided at a randomly selected crossover point, and the resulting sub-trees are exchanged to form two new individuals (model structures). Before calculation of model parameters using the least squares method the cut procedure is provided for deleting identical model sub-trees (regressors). The model selection is stopped when the fitness value ϕ_j exceeds the target value, or the maximal number of generation (iterations) is reached.

Speaking less technically, the process of regressors selection according to evolutionary principle is similar to the biological process of organisms survival. In the same way as in nature, genetic algorithms search perfect individuals without using information about them. Every individual has a fitness value that expresses the efficiency of the corresponding solution for describing the data. Better solutions are assigned higher values of fitness than worse solutions. The fitness function also determines how successful the individual will be at propagating its genes to subsequent generations in the next population. It permits to select the most adapted individuals according to the evolution survival principal. After that the fitness functions values are calculated and the most enduring individuals are chosen to generate the new population by randomly applying one of genetic operators: mutation, recombination and crossover.

The implementation of this procedure used in PROGRESS was developed by Semeniv (2015). It uses individual fitness functions in the form of linear correlation with the pre-

dictand, with a penalty for model complexity. The main difference with earlier instances is the degree of polynomial nonlinearity: the article Semeniv (2015) was limited to linear and bilinear models, while in PROGRESS fifth power polynomial models are used.

3.2.3 Regression modelling approach

The Regression Modelling (RM) is a method for constructing polynomial NARX models of strongly nonlinear stochastic dynamical systems with feedback, featuring processes with vastly different time scales, based on ANOVA. It was developed by SRI NASU-NSAU in the context of space weather prediction (Parnowski 2011). It constructs the model over a large enough training sample in several steps: first an AR model is constructed (in several smaller steps for performance reasons), then an ARX model, and finally a NARX model. At each step insignificant regressors are identified with the F-test and discarded; the procedure repeats until only significant regressors remain. The parameters are determined using the least squares method. Nonlinear regressors are constructed from the most significant regressors of the ARX model. After the model is constructed, it is once again put through the F-test, but on a different sample to reduce overfitting. The resulting model stays current for at least one solar cycle unlike some other approaches, which require the model to be updated annually. Of course, this leads to greater complexity.

The difference to the models developed earlier using the same approach is the greater degree of polynomial nonlinearity: 4 instead of 3.

3.2.4 FROLS and the ERR

In contrast to the NARMAX methodologies mentioned above in Section 3.2.2 and Section 3.2.3 that have been developed at SRI, the NARMAX implementation developed at USFD uses a different set of methods to determine the model structure.

The USFD NARMAX methodology employs the Forward Regression Orthogonal Least Squares (FROLS) algorithm and its several variants (Billings et al. 1989, Chen et al. 1989, Wei et al. 2004, Wei & Billings 2008, Billings 2013) to fit a NARMAX model based on the input and output data sets by identifying the monomial terms that have the greatest influence on the evolution of the model output parameter data set. Candidate monomial terms are calculated using the mathematical combination of the input parameter terms using a predefined set of mathematical operators (such as $\{+, -, \times, \div, \sqrt{}\}$). The candidate terms may be composed of linear, quadratic, or higher terms of the input parameters measured at a set of discrete time lags. This process may result in many thousands of possible candidate terms, most of which have very little influence on the system output.

Within FROLS, the selection of the most significant terms is made based on the calculation of the Error Reduction Ratio (ERR). The ERR quantifies the proportion of the variance of the output signal that may be attributed to each candidate monomial term, ranking them from highest to lowest. The term with the highest ERR value is chosen as a model term and its contribution to the system output is taken into account using an orthogonalisation procedure. This process is then repeated to reveal other model terms until either sufficient terms have been extracted to account for the majority of the input signal variance or the model residual is purely random.

USFD is currently developing a Machine Learning Enhanced NARMAX Model Averaging framework and a NARMAX Model Ensemble approach, implemented through modern sparse dictionary learning techniques, to enhance and improve models' robustness and reliability.

Once the set of model terms has been identified their coefficients are determined using a least squares fitting algorithm.

3.2.5 Bi-linear NARMAX models

As briefly mentioned in Section 3.2, and elaborated on further in Section 3.2.4, the NARMAX methodology is capable of approximating the output signal of a system by determining a set of model terms based on the input parameters of the system. In the general NARMAX methodology, all linear, quadratic, and higher power input terms may be considered in virtually any combination to deduce the best model.

Bi-linear models represent a subset of NARMAX models. This class of models are characterised by being composed of only the set of cross product terms of explanatory variables that may be described as 'linear in the parameter'. As an example, consider a system with two sets of input scalar values, x, and y. A bilinear model is simply the weighted sum of combinations of these parameters.

$$\mathbf{z} = \sum_{i,j} \mathbf{m_{i,j}} x_i y_j \tag{2}$$

4 Data sets used to develop the models

The models developed here are based on the analysis of near real time (NRT) data sets available from the OMNI2 database, maintained by NASA GSFC, SPDF and NSSDC. The parameters representing the system input are IMF (total intensity, 2 angular and 3 Cartesian components) and SW plasma parameters (density, proton temperature, and velocity), together with previous values of the system output, either the *Dst* or *Kp* index. Typically, values with an hourly cadence are used as measurements of the input and output parameters of the system.

5 Validation techniques

The validation of GNM is somewhat problematic due to its output being an interval, and not a single value. A quick analysis of existing validation techniques showed that there are no readily available solutions for this case.

The naïve approach to validation of interval forecast would be measuring the percentage of cases when the observed value was within the prediction interval. Ideally, it should match the confidence level of the interval, but in reality it is slightly different due to different statistical properties of training and validation samples. It naturally prefers forecasts with large prediction intervals, which have little practical value, so its usefulness is rather objectionable, however we still calculated this score.

Fortunately, it is possible to treat this type of forecast as a dichotomous, a probabilistic, or a multi-category (in case of *Kp* index) forecast (see http://cawcr.gov.au/projects/verification/ for details).

The most productive approach is treating the interval forecast as a probabilistic forecast using the following convention:

- the storm is defined as $Dst \leq$ -50 nT for the *Dst* forecast and $Kp \geq 50$ ($Kp \times 10 \geq 50$) for the *Kp* forecast;
- if the guaranteed interval is defined not as a confidence interval, it is assigned a 90% confidence level;
- if the whole interval is in the storm region, the probability of the storm is defined as $pi = (1 + CL)/2$, where CL is the confidence level of the interval;
- if the whole interval is in the calm region, the probability of the storm is defined as $pi = (1 - CL)/2$;
- if the interval crosses the boundary between the calm and the storm regions, the probability of the storm is defined as $pi = (1 - CL)/2 + CL \times q$, where q is the fraction of the interval?s length located in the storm region.

The last item is justified by the Bayes axiom due to the requirement of GE approach not to assume any specific distribution of errors.

Another related problem is that there is no evident reference forecast to compare to, because it seems impossible to define either persistence or climatology interval forecasts. Thus, for a two-category (storm or calm) probabilistic forecast the only relevant metric is the Brier score, defined as:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2 \tag{3}$$

where o_i is the indicator if a storm actually occurred (1 is storm, 0 is calm), and N is the number of data points. The Brier score ranges from 0 to 1 with 0 being a perfect score. Note that it is favourable towards rare events, so it can be used to compare different forecasts only over samples with equal storm occurrence rates. Still, this is probably not the perfect approach, as it completely ignores the actual observed value.

The dichotomous treatment is similar to the probabilistic one with the confidence level set to 100%, and the linear dependence on q in the last item replaced with a Heaviside step function going from 0 to 1 at $q = 0.5$. Then, a contingency table can be constructed together with all the familiar metrics based on it. However, this approach does not depend on the length of the prediction interval, and is the same as for a deterministic forecast, so we do not use it.

It is also possible to treat guaranteed *Kp* forecast as a multi-category forecast following the same guidelines. In this case, ranked probability score should be used instead of the Brier score. We did not pursue this validation strategy to retain comparability between *Dst* and *Kp* forecasts.

Validation of AE was performed in the same manner as for Dst and Kp models. The boundary value of the AE index for the calculation of Brier score was 400 nT.

A program for calculating the Brier score is available in the form of FORTRAN source code with a configuration file.

6 Artificial Neural Network Models

Forecast models for the geomagnetic indices Kp Dst and AE have been developed by IRF based on ANN methodologies. These models, known as IRF-Kp-2017, IRF-Dst-2017, and IRF-AE-2017,

The IRF-Kp-2017 model is an ensemble of ANNs to predict the Kp index. The inputs to the IRF-Kp-2017 model are solar wind plasma and magnetic fields, and periodic functions of time-of-year and time-of-day. The high resolution ACE L2 data are first propagated from L1 to upstream the Earth and then transformed to 3 hour resolution using averages, minima, and maxima of the solar wind parameters. The model predicts with lead times of 0, 1, 2, and 3 hours in excess of the propagation lead time. Further details may be found in PROGRESS deliverable D3.4.

The IRF-Dst-2017 model is an ensemble of ANNs to predict the Dst index. The inputs to the IRF-Dst-2017 model are solar wind plasma and magnetic fields, and periodic functions of time-of-year and time-of-day, using 1-hour averages. The model was developed using OMNI data, therefore no propagation was required. However, for real-time operation or testing using ACE L2 data the solar wind must first be propagated. The model predicts with lead times of 0, 1, 2, and 3 hours in excess of the propagation lead time. Further details may be found in PROGRESS deliverable D3.4.

7 Guaranteed NARMAX Models

7.1 Data source

Development of the Kp, Dst, and AE GNM is based on data from the OMNI2 database, maintained by NASA GSFC, SPDF and NSSDC.

For the Kp (SRI-Kp-GP, SRI-Kp-RM) and Dst (SRI-Dst-GP, SRI-Dst-RM) data with an hourly cadence were selected. Naturally, only those parameters which are available in NRT, namely the IMF (total intensity, 2 angular and 3 Cartesian components) and SW plasma parameters (density, proton temperature, and velocity), and previous values of either Dst or Kp index were used.

For construction of GNM with GP algorithm we used the annual dataset for 2006 for training, from which only the product $V \times B_z$ and the previous values of the predictand with a maximum lag of 54 hours were used.

For construction of GNM with RM algorithm we used 3 samples: years 1976 to 2000 for model training, and 2001 to 2008 for model tuning (these two have approximately equal number of data points). We did not use the data from 2009 due to anomalously quiet solar wind conditions at that time. All IMF and SW plasma data with lags up to 24 hours, as well as previous values of the predictand with lags up to 27 days (1 Carrington

period of the Sun) were used. Using large maximum lag for geomagnetic indices allows our models to take into account recurrent space weather events. To simulate diurnal and seasonal variations we also added 4 synthetic inputs, which are simply sine and cosine functions with periods of 12 hours and 6 months. Together with either of two geomagnetic indices this makes 14 inputs.

For the validation of both versions of GNM we used a sample from 2010 to 2017 and annual samples for years 2014 through 2017.

Development of the *AE* GNM is based on hourly cadence data from the OMNI2 database for the period from 2013-03-12T11:00:00Z to 2013-06-03T18:00:00Z (2000 hours) for training, from which only the product $V \times B_z$ and the previous values of the predictand without lag were used.

As in the case of the *Kp* and *Dst* models, the *AE* model validated using annual samples for years 2014 through 2017.

7.2 Implementation

Using the GNM methodology, several forecast models have been generated. As well as generating models using the GP and RM algorithms, the forecast horizons were also varied. The list of models is shown in Table 1.

A single GNM model for the *AE*was built using a GP algorithm for structural identification and least squares method for parametric identification. A complete description is given in Section 3.2.1.

In the GP version of GNM, the half-width of the prediction interval is defined as the maximum difference between adjacent values of the predictand in the training sample.

The GP version of GNM is available as a set of MATLAB .m files, which contain the models for both *Dst* and *Kp*ṪThe user can choose an input data file in OMNI2 format, the index to be predicted, and the number of records to be processed in the input file. The output is provided as a plot and as an ASCII file, which contains the number of the record, the observed value of the index, and the bottom and top boundaries of the prediction interval. The format string for the *Kp* forecast is (I5, 4I6), the *Dst* forecast (I6, F7.0, 2F9.2), and the *AE* (6E16.7). Basic documentation is provided in the form of a readme file.

In the RM version of GNM, the half-width of the prediction interval is defined as two root mean square errors measured on the test sample, so the guaranteed interval is a roughly 95% confidence interval.

The RM version of GNM is available as a set of source code files adhering to FORTRAN 90 fixed form standard (.for), a set of ASCII files containing the models (.res, .cov), and a set of ASCII configuration files (.cfg). The programs are controlled by configuration files, which contain basic usage instructions. The user can choose which parameter to forecast, with what lead time, set input data file name, format, and fill values, set usage flags, cadences, and maximum lags per parameter, choose which model to use, the name and the format of the output file, and the name of the metadata file. Models (.res) are written in a format readable both by a machine and a human, which allows notating an arbitrary polynomial. Each model is supplied with a covariance matrix (.cov) used to calculate a guaranteed interval. The output file contains at most the following data: the number

of the record, year, month, day of month, day of year, UT hour, the observed value of the index, and the bottom and top boundaries of the prediction interval. Only ASCII output is provided; plots should be produced by the user. The metadata file contains the following information: number of the predictand, lead time, cadence of the predictand, start and stop times of the output file, root mean square error, prediction efficiency, and linear correlation coefficient of the prediction baseline and of the persistence with respect to observation, and the skill score (relative reduction of root mean square error in comparison to persistence). Basic documentation (readme file) is provided.

8 Bilinear models

8.1 Data source

The data with which the NARAMX bilinear models are build comes from the OMNI2 database, the same as that used for the GNM model mentioned in Section 7.1.

For the bilinear models, data with a cadence of 1 minute were collected and averaged over a 1 hour period to obtain the input data sources for the models. Any data gaps were filled using the previous measured value.

8.2 Implementation

The models were created using the FROLS and ERR methodology, described in Section 3.2.5, to select the most important model terms and their coefficients. These models were them implemented using two different methodologies to produce one step ahead (OSA) forecasts, and a model predicted output (MPO).

OSA forecasts are generated by taking the set of measured input parameters i.e. measurements of the solar wind and previous values of the Kp index and using them to estimate the next value for Kp as described in Equation 4. This process is repeated to obtain the next forecast estimate of Kp. Since the output from the OSA implementation is completely based on the availability of measurements the forecast horizon is limited (one time step) but their accuracy remains high.

$$\hat{y}(t) = F[u(t-1), ..., y(t-1), ...] \tag{4}$$

In contrast, MPO forecasts use previous forecasts of the output parameter. These models are typically primed with a few initial measurements of the input parameters and then use previous forecasts within their set of input parameters (as described in Equation 5).

$$\hat{y}(t) = F[u(t-1), ..., \hat{y}(t-1), ...] \tag{5}$$

The values of the Kp index output from these models have not been artificially constrained to lie within the range $0 < Kp < 9$. As a result, forecast values greater than 9 are possible.

9 Model Results and Discussion

9.1 GNM models

The GNMs for the three geomagnetic indices were validated using annual datasets from year 2014 to year 2017. The results are given in Tables 2 and 3. Table 2 shows the percentage of correct forecasts. A correct forecast occurs when the measured value of the index lies between the upper and lower confidence limits as defined by the model. The higher the percentage score, the better the model forecast. Table 3 contains the Brier score for all datasets.

Table 2: Percentage of correct forecasts for the GNM models.

Model	Years				
	2014	2015	2016	2017	2010-2017
GP *Dst* 1h	97.4%	97.1%	98.1%	92.7%	91.8%
GP *Kp* 3h	86.1%	88.3%	85.9%	65.9%	67.9%
GP *AE*	94.8%	91.4%	91.9%	92.2%	
RM *Dst* 1h	96.3%	96.4%	96.4%	96.2%	96.3%
RM *Dst* 3h	96.0%	96.3%	95.8%	96.2%	96.0%
RM *Kp* 3h	95.9%	96.9%	95.6%	97.4%	95.5%

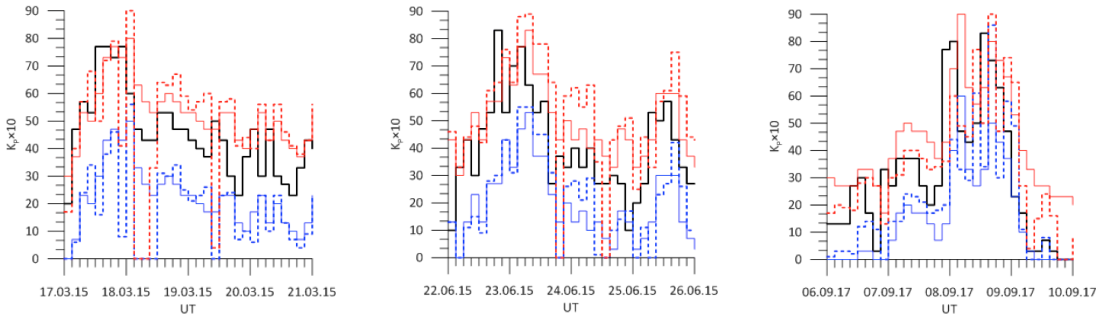
Table 3: Brier scores for the GNM models.

Model	Years				
	2014	2015	2016	2017	2010-2017
GP *Dst* 1h	1.21%	2.02%	0.97%	0.77%	1.22%
GP *Kp* 3h	0.25%	0.25%	0.25%	0.25%	0.25%
GP *AE*	15.69%	19.37%	18.87%	16.51%	
RM *Dst* 1h	0.41%	0.92%	0.48%	0.43%	0.51%
RM *Dst* 3h	0.97%	1.99%	1.01%	0.75%	1.06%
RM *Kp* 3h	2.48%	9.44%	5.75%	8.32%	3.93%

As an example of the forecasts generated by the GNMs, Figures 1, 2, and 3 show the upper (blue) and lower (red) edges of the 95% confidence level interval for the forecasts of *Kp*, *Dst*, and *AE*. The black line represents the calculated values of the index based on observations.

Figure 3 shows the output of the GNM model for the storms observed on March 17-18, 2015 (panel a), June 22-23, 2015 (panel b), and September 7-8, 2017 (panel c).

The model for the *AE* index performed worse than those for *Dst* and *Kp*. This can be caused by the following reasons:



(a) March 17-18, 2015 (b) June 22-23, 2015 (c) September 7-8, 2017

Figure 1: Retrospective GNM forecasts for the variation in the *Kp* index. Thin solid blue and red lines are low and high edges of the RM interval with 3h lead time, thick dotted red and blue lines are low and high edges of the GP interval with 3h lead time.

- The *AE* index has questionable physical meaning and is plagued by contaminations from regular variations S_R, subsurface currents, latitude gap between eastward and westward auroral electrojets, and uneven distribution of contributing stations. This was first noted by Rostoker (1972), and a detailed discussion was given by Mayaud (1980) with some additional points made by Kamide & Rostoker (2004).

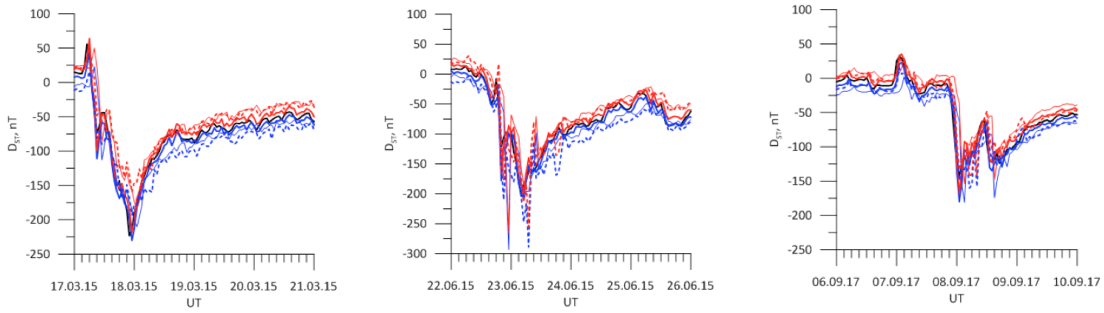
- A major contribution to the *AE* index is provided by substorm activity, which is caused by poorly understood physical processes in the magnetotail, which seem to be only partially dependent on solar wind conditions. Thus, L1 data are not sufficient to correctly describe its dynamics.

- The *AE* index manifests chaotic dynamics and has higher first Lyapunov exponent than *Dst* and *Kp* so it is less predictable. It could potentially be better forecast with a chaotic predictor, but there are no readily available methods to construct one other than by trial and error.

9.2 Bi-linear models

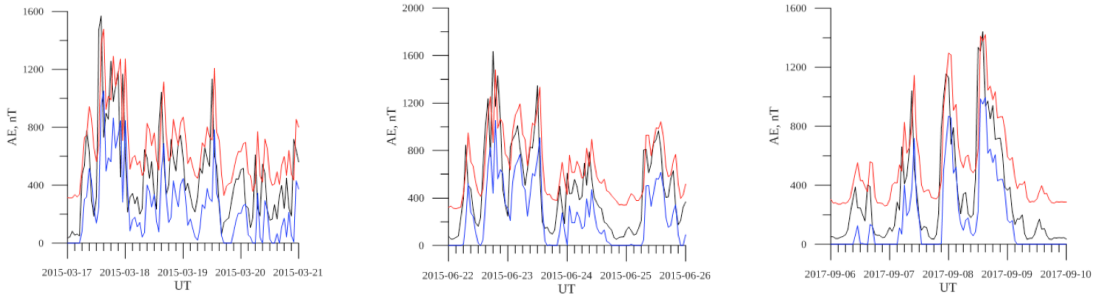
Examples of the forecasts of the NARMAX bi-linear models for the *Kp* and *Dst* are shown in Figure 4 and Figure 5 respectively for the three geomagnetic storm periods shown in Figures 1, 2, and 3. Figure 4 shows the measurements of *Kp* in blue and the OSA and MPO model output in red.

Figure 5 Shows the forecasts for the three geomagnetic storm observed around March 17, 2015 (panel (a)), June 22, 2015 (panel (b)), and September 8, 2017 (panel (c)). The black line represents measurements of *Dst* index. The red and green lines show the forecasts generated by the one and three hour ahead forecast models while the blue line shows MPO results based on the the model that was initialised at the start of the month and left to run based on previous forecasts. On the whole, the one and three hour ahead models do capture the overall variation of *Dst* during the storms, exhibiting similar onset and decay times to those observed. The MPO implementation tends to correlate less



(a) March 17-18, 2015 (b) June 22-23, 2015 (c) September 7-8, 2017

Figure 2: Retrospective GNM forecasts for the variation in the *Dst* index . Thick solid black line is observation, thick solid blue and red lines are low and high edges of the RM interval with 1h lead time, thin solid blue and red lines are low and high edges of the RM interval with 3h lead time, thick dotted red and blue lines are low and high edges of the GP interval with 1h lead time.



(a) March 17-18, 2015 (b) June 22-23, 2015 (c) September 7-8, 2017

Figure 3: Retrospective GNM forecasts for the variation in the *AE* index . The black line is observation, blue and red lines are the low and high edges of the prediction interval with 1h lead time.

well with measured values however, it does tend to capture the storm onset times fairly accurately.

Performance statistics for the two *Dst* models are shown in Table 4. While most statistics point to a good models performance, it should be noted that they are out performed by full NARMAX models since bi-linear models may not be sufficient to characterise the processes occurring. It can be seen from Figure 5 that while the models were able to forecast the event on March 17-18 2015 to a high degree, the models performed less well on the other two events considered. This may be an artefact of the use of the shorter training period (two years data) than was the case with the GNM models.

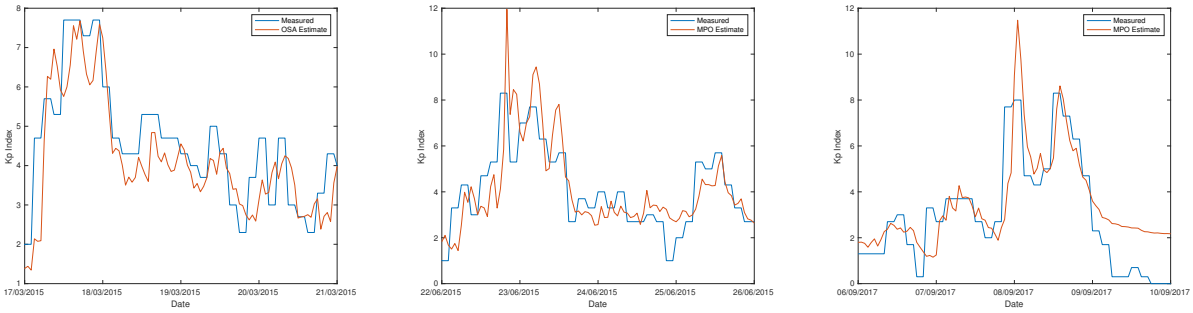
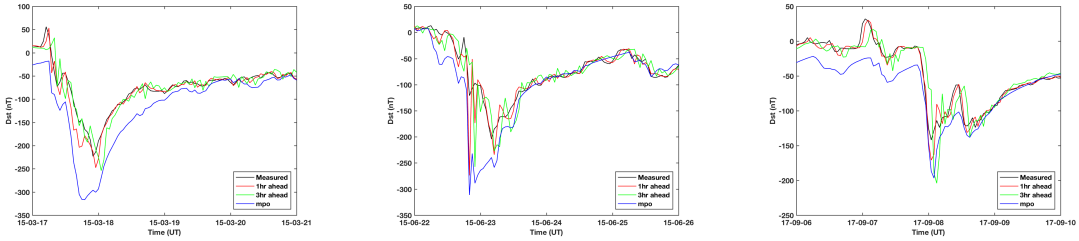


Figure 4: Retrospective forecasts of the *Kp* index for the major geomagnetic storms that occurred around March 17, 2015 (left), June 22, 2015 (centre), and September 8, 2017 (right). The blue lines show the measured values of *Kp*, while the red show the estimates of the bilinear One Step Ahead and Model Predicted Output models.



(a) March 17-18, 2015 (b) June 22-23, 2015 (c) September 7-8, 2017

Figure 5: Retrospective bilinear forecasts for the variation in the *Dst* index for the three geomagnetic disturbances periods shown in Figures 1, 2, and 3.

10 Lyapunov Exponents of the *Dst* index

When performing forecasts, it is important to investigate the time horizon beyond which the prediction interval is large enough to render any forecasts useless. From a dynamical systems perspective the rate at which forecasts of the system diverge from actual system measurements can be investigated by calculating the Lyapunov exponents of the system.

Chaotic systems are a class of complex dynamic systems whose motions are extremely sensitive to initial conditions (Sprott 2003). Trajectories starting with two very close initial values will normally separate from each other at an exponential rate over time. Lyapunov exponents were developed to quantitatively measure a nonlinear dynamic system's chaotic property by evaluating the separation behaviour (contraction and expansion in different directions in phase space) of two orbits which are initially very close to each other. Lyapunov exponents do not measure transient local behaviour of a system, but an overall behaviour of the system through a long period of evolution.

In the Lyapunov spectrum, the smallest Lyapunov exponent characterises the speed of convergence where as the largest Lyapunov exponent signifies the speed of divergence. In practice, the value of the largest Lyapunov exponent (LLE) plays an important role, this

Table 4: Forecast statistics for the bi-linear *Dst* models with 1 and 3 hour ahead predictions.

Dst **model**	**Period**	**Corr**	**PE**	**MSE**	**RMSE**	**NRMSE**
1hr	March 2015	0.9807	0.9543	49.8087	7.0575	0.2137
	17-21 March 2015	0.9698	0.9064	175.8363	13.2603	0.3060
3hr	March 2015	0.9422	0.8812	130.5901	11.4276	0.3447
	17-21 March 2015	0.9320	0.8193	313.3827	17.7026	0.4251
1hr	June 2015	0.9662	0.9277	72.3230	8.5043	0.2689
	22-26 June 2015	0.9089	0.7872	346.4754	18.6135	0.4613
3hr	June 2016	0.8529	0.6400	580.6235	24.0961	0.6000
	22-26 June 2015	0.8520	0.6379	584.4763	24.1759	0.6018
1hr	Sept 2017	0.9775	0.9543	26.0482	5.1037	0.2139
	6-10 Sept 2017	0.9808	0.9608	59.1425	7.6904	0.1980
3hr	Sept 2017	0.8823	0.7618	128.9672	11.3564	0.4880
	6-10 Sept 2017	0.8813	0.7516	372.3170	19.2955	0.4984

is because if the largest LLE is positive, then it means that the system is chaotic; if LLE is equal to zero, it then indicates that there exist periodic or quasi-periodic dynamics in the process (Eckmann & Ruell 1985).

In the present study of the Lyapunov exponents of the *Dst* index, the algorithms developed in Gencay & Dechert (1992)) and Lai & Chen (1998) were used to calculate Lyapunov exponents of *Dst* index data for a period of 16 years. The largest and smallest values of Lyapunov exponents for years 1998 to 2014 are shown in Figure 6. It can be seen that both the largest and smallest Lyapunov exponents of *Dst* index are negative, meaning that *Dst* index does not show any chaotic behaviour. Our results further confirm the argument and conclusion given in Temerin & Li (2002) that "the magnetosphere is highly predictable and that chaotic behavior within the magnetosphere has little influence on the large-scale currents that determine Dst".

11 Model comparison

11.1 Input and target data sources

All models use solar wind data for the inputs and we base the analysis on the ACE L2 plasma and magnetic field 64 second data and the OMNI 1-hour resolution dataset. The models assume that the solar wind measurements are taken at a location close to the Earth's bow shock. As ACE measures at a location around L1 each sample is temporally shifted from spacecraft location to close to Earth before being temporally transformed. We simply assume that the travel time of a solar wind sample is $(x_{L1} - x_{BS})/V$ where we set $x_{BS} = 10\mathrm{R_E}$ and V is the solar wind speed.

For the target data we use *Kp* index from GFZ, *Dst* and *AE* indices from WDC-Kyoto. The temporal resolution of *Kp* is 3 hours, *Dst* 1 hour, and *AE* 1 minute. No further processing is applied on *Kp* or *Dst* but *AE* is temporally averaged to fit the

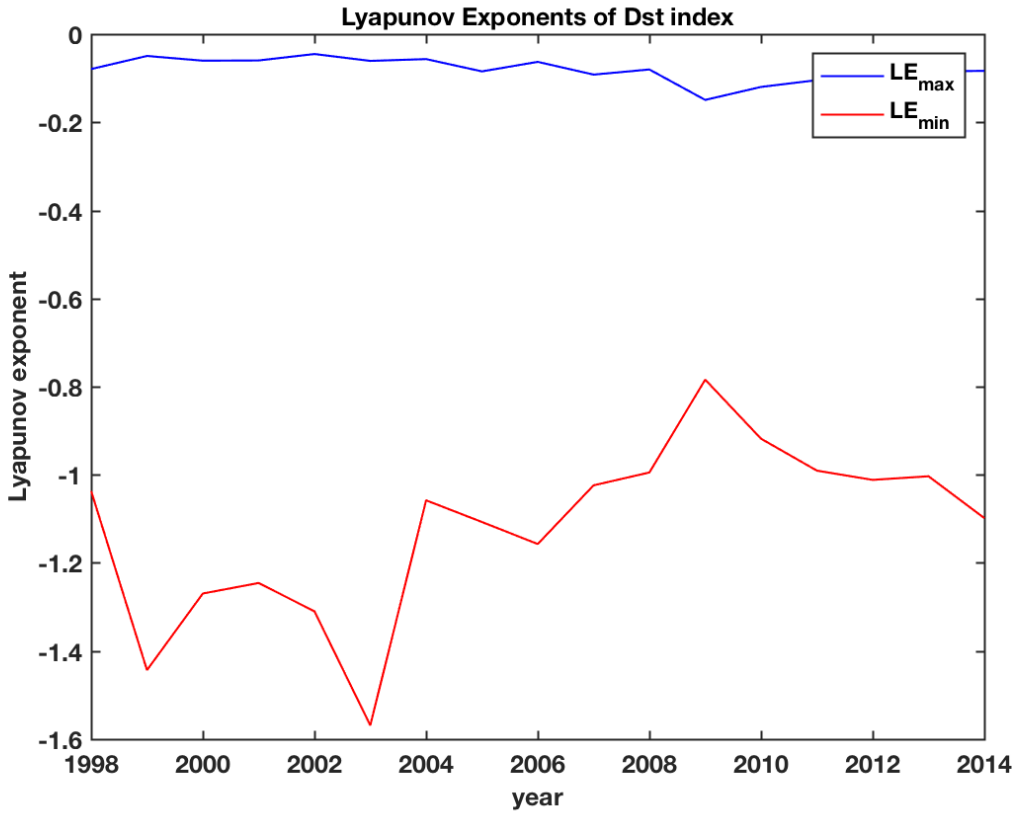


Figure 6: Variation in the annual maximal and minimal Lyaponov exponents ofthe *Dst* index for the period 1998 to 2014.

resolution of the models.

The number of *Kp* values in different intervals per year is shown in Figure 7. Top panel shows all available *Kp* values, while the bottom panel only counts those *Kp* values that have simultaneous ACE 3-hour average data. For example, in 2009 all *Kp* values are less than 6.

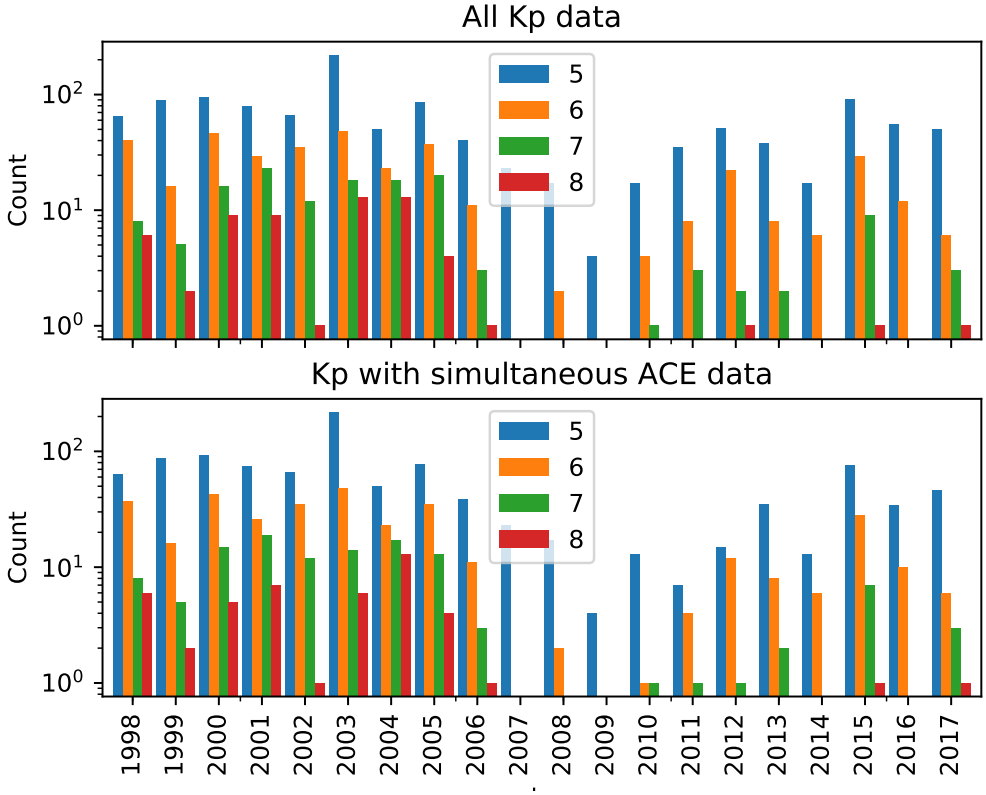


Figure 7: Number of *Kp* values for different intervals per year. Level 5 corresponds to $5 \leq Kp < 6$ and so on up to level 8 with $Kp \geq 8$.

Similarly, the number of *Dst* values in different intervals per year is shown in Figure 8. Top panel shows all available *Dst* values, while the bottom panel only counts those *Dst* values that have simultaneous OMNI 1-hour average data. In 2009 all *Dst* values are above -100 nT.

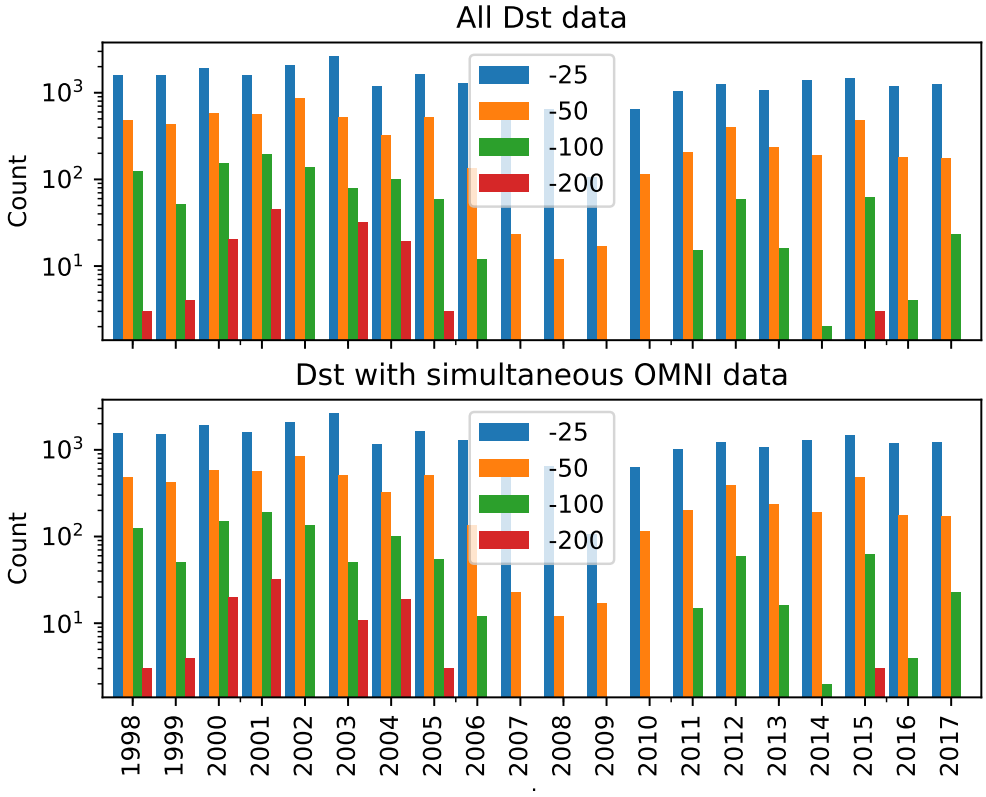


Figure 8: Number of *Dst* values for different intervals per year. Level -25 corresponds to $-50 < Dst \leq -25$ and so on up to level -200 with $Dst < -200$.

11.2 Statistical comparisons

We apply the following statistics on the prediction models:

$$\text{BIAS} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i - y_i \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - y_i)^2} \tag{7}$$

$$\text{CORR} = \frac{\sum_{i=1}^{n}(\bar{y}_i - \langle\bar{y}\rangle)(y_i - \langle y\rangle)}{\sqrt{\sum_{i=1}^{n}(\bar{y}_i - \langle\bar{y}\rangle)^2}\sqrt{\sum_{i=1}^{n}(y_i - \langle y\rangle)^2}} \tag{8}$$

$$\text{R2} = 1 - \frac{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \langle y\rangle)^2} \tag{9}$$

where $\bar{y}$ is predicted index, y is observed index, and $\langle\cdot\rangle$ is the mean. For a perfect model we have BIAS=0, RMSE=0, CORR=1, and R2=1. In the case of the GNM models, the above scores were calculated for the baseline NARMAX model. Scores that take into account the interval nature of these forecasts were given in Section 9.1.

Table 5 summarises the measures for all *Kp* models over all data 1998–2017. It should

be noted that the statistics have been computed on data that to different degrees contain training data, see Table 1. For comparison we also include the persistence model (Pers) which simply is

$$\bar{y}(t) = y(t - \Delta t) \tag{10}$$

where Δt is the temporal resolution. Thus, predicted value is simply the same as the past value. All models have BIAS close to zero. Considering RMSE, CORR, and R2 the IRF-Kp-2017-T0 and -T1 models scores best and have very similar measures. On 3-hour lead time the IRF-Kp-2017-T3 and SRI-Kp-RM-T3 are the best, with the former slightly better. All models except the SRI-Kp-GP-T3 scores better than persistence.

Table 5: Statistical measures for all data in 1998–2017 for the *Kp* prediction models. Tn indicates lead time with n in hours.

	BIAS	RMSE	CORR	R2
Pers	0.00	0.85	0.81	0.62
IRF-Kp-T0	-0.04	0.53	0.92	0.85
IRF-Kp-T1	-0.03	0.53	0.92	0.85
IRF-Kp-T2	-0.02	0.61	0.90	0.80
IRF-Kp-T3	0.01	0.72	0.85	0.73
USFD-Kp-OSA-T1	0.10	0.71	0.86	0.73
USFD-Kp-MPO-T1	0.20	0.78	0.84	0.68
SRI-Kp-RM-T3	-0.07	0.81	0.82	0.66
SRI-Kp-GP-T3	-0.04	1.17	0.62	0.29

Table 6 summarises the measures for all *Dst* models also over all data 1998–2017. The shorter lead-time models (T0 and T1) always scores better than the 3-hour lead-time (T3) models. The BIAS is generally very small. The SRI-Dst-RM-T1 scores best closely followed by persistence.

As the models have been derived from data that to different degrees are part of the data used for the statistics in Tables 5 and 6 it is useful to study the measures for each each. Figure 9 shows the correlation (CORR) for each model and year, with years not included in the training set marked with dots. There is some variation over the years, but the correlation computed on years with training data do not show systematically higher values, indicating that the measures in Tables 5 and 6 are valid. It is also interesting to see that 2009 shows lower correlations for all models.

Figure 10 shows RMSE as function of year. For most models the minimum RMSE occurs in 2009, partly an effect of the very low levels of activity.

In Figure 11 the linear correlations for the *Dst* models are shown. The SRI-Dst-RM-T1 model consistently shows highest correlation, and higher that persistence except for 2017. The SRI-Dst-GP-T1 also shows high correlation, but not as consistently, several years before 2006 have lower correlations. For several of the models 2009 again stands out with poor performance.

Table 6: Statistical measures for all data in 1998–2017 for the *Dst* prediction models. Tn indicates lead time with n in hours. BIAS and RMSE are given in nT.

	BIAS	RMSE	CORR	R2
Pers	0.01	4.40	0.98	0.95
IRF-Dst-T0	0.69	8.70	0.91	0.82
IRF-Dst-T1	-0.15	8.80	0.90	0.82
IRF-Dst-T2	0.07	9.31	0.89	0.80
IRF-Dst-T3	-1.19	10.04	0.88	0.76
USFD-Dst-BL-T1	0.44	4.48	0.98	0.95
USFD-Dst-BL-T3	0.13	8.10	0.92	0.85
SRI-Dst-RM-T1	-0.01	3.27	0.99	0.97
SRI-Dst-RM-T3	0.07	7.24	0.94	0.88
SRI-Dst-GP-T1	-0.10	6.10	0.96	0.91

The RMSE for the *Dst* models are shown in Figure 12.

12 Discussion and Conclusions

We have analysed the performance of different index prediction models driven by upstream solar wind for the years 1998 to 2017. We now discuss the results.

Regarding the inputs the models differ in two fundamentally different ways: with or without the past target values. As the autocorrelations of the indices are strong that means that providing past values will in a statistical sense improve the predictions. For *Kp* the one- and two-step autocorrelations are 0.81 and 0.69, respectively, corresponding to 3 and 6 hours lag. For *Dst* it is even stronger with 0.98, 0.94, and 0.90 for 1-, 2-, and 3-step lags, respectively. There are several reasons for the lower autocorrelation seen in *Kp* compared to *Dst*, but one aspect is that the *Kp* index is a range index and effectively filters out low frequency geomagnetic variation improving on its stationarity.

There is a monotonic decrease in performance for the IRF-Kp-2017 model, which do not use past *Kp* as inputs, when going from the 0 and 1 hour models, to the 2, and 3 hour models (Table 5 and Figure 9). This is understood in terms of that the only available lead-time from a solar wind monitor at Earth bow shock location is up to 1 hour, after which the predictions start lag. As the SRI-Kp-RM model only provides 3-hour lead-time predictions it basically picks up the autocorrelation in the *Kp* series. Thus, past *Kp* values are not crucial to the performance.

For the *Dst* models the situation is quite different. Mapping from solar wind only (IRF-Dst-2017 model) is more difficult possibly due to changes in baselines (*Dst* quiet time levels) and including past observed *Dst* will help. The autocorrelation is also strong,

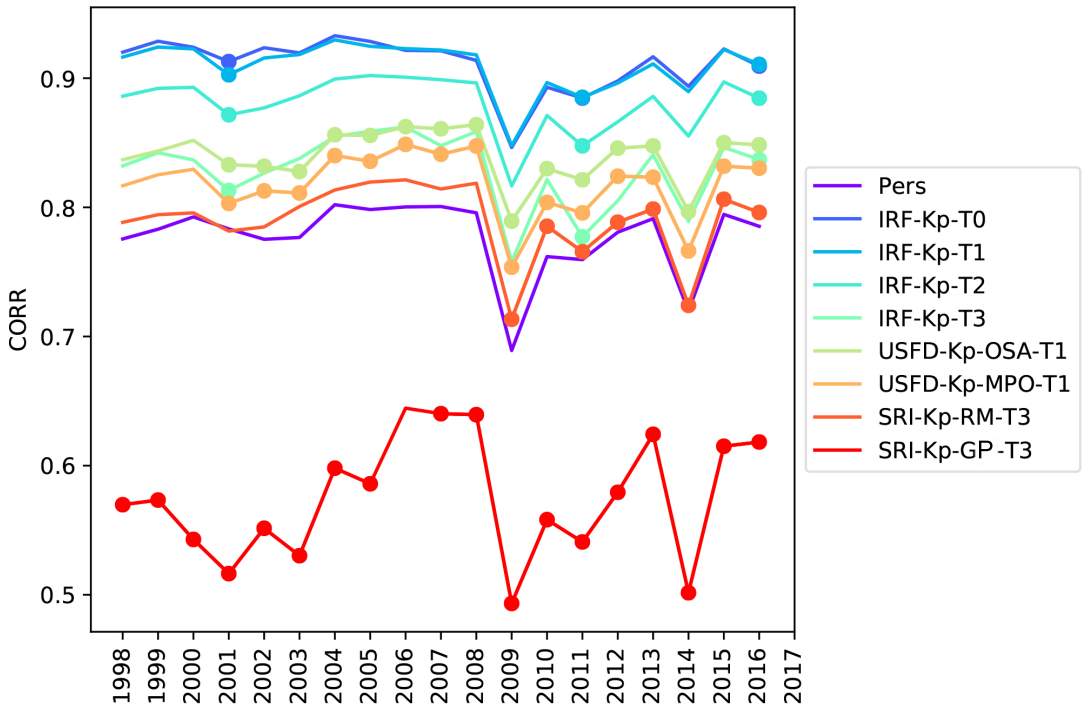


Figure 9: Linear correlation between predicted and observed *Kp* for the different models as function of year. Dots indicate that the year was not part of the training set.

partly attributed to the long decay times of the recovery phases. However, at the onset of the storms (main phases) the lead time is limited to about 1 hour. The SRI-Dst-RM-T1 model (Figure 11) generally show higher correlation than persistence, except for 2017, as it utilises both the strong 1-hour autocorrelation and the driving solar wind. However, at 3-hour lead-time (SRI-Dst-RM-T3) the solar wind does not provide predictive capabilities and the predictions becomes similar to persistence.

By differentiating the *Dst* index and its predictions we can effectively remove any non-stationarity and varying base-lines, and also focus on changes in *Dst*. By selecting only data for which *Dst* is below its average ($Dst < -13$ nT) and when its derivative is negative ($dDst/dt < 0$) we focus on the *Dst* main phases. Figure 13 shows the linear correlation. The correlation has dropped for all models, as expected, but now the IRF-Dst-T0 and -T1 models are above persistence.

When the models are operated in real time there always issues with delays in the different subsystems of the prediction models. The models used here are computationally lightweight which means that they introduce insignificant delays, the main delays come from measurements and their distribution. Any delays in this chain will reduce the true prediction lead time. Under normal conditions the delay of the SWPC solar wind data typically lies between 3 to 5 minutes. The construction of the geomagnetic indices from measurements show longer delays which must be considered if they are used as inputs

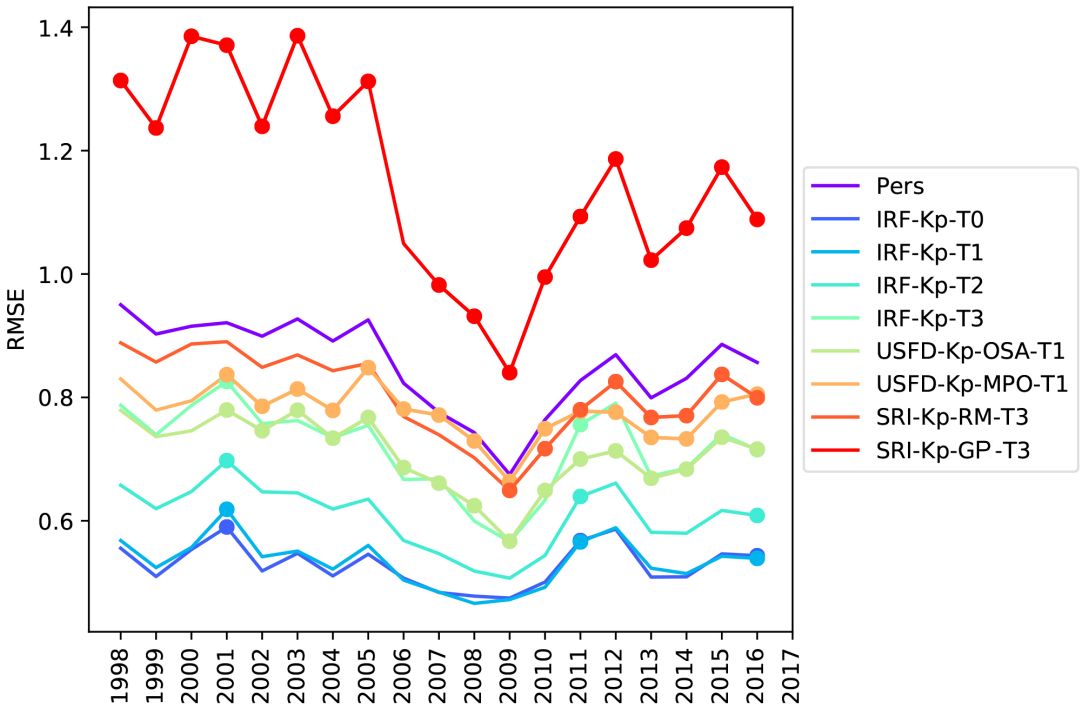


Figure 10: RMSE between predicted and observed *Kp* for the different models as function of year. Dots indicate that the year was not part of the training set.

to prediction models. GFZ provides a first estimate of the *Kp* index about 1.5 hours into the latest 3 hour interval, known as nowcast *Kp*The latest *Kp* value is then updated until typically 30 minutes after the interval has finished. The real-time *Dst* index from WDC-Kyoto is typically published 30 minutes into the latest 1-hour interval and may be updated for 30 minutes after the end of the interval. For both *Kp* and *Dst* further changes may also occur later as better estimates of baselines are determined. For *AE* there are no publicly available real-time data.

To conclude:

- Guaranteed NARMAX Models for *Dst*, *Kp* and *AE* indices were constructed using two different algorithms by SRI.
- All the developed forecasts with the exception of the GP-based Kp forecast, which has a too wide prediction interval, provide useful information and are ready for transition to near-real time operations.
- Bi-linear models for the *Kp* and *Dst* indices were constructed by USFD.
- The performance of each of the models generated within PROGRESS were assessed for common periods of data.

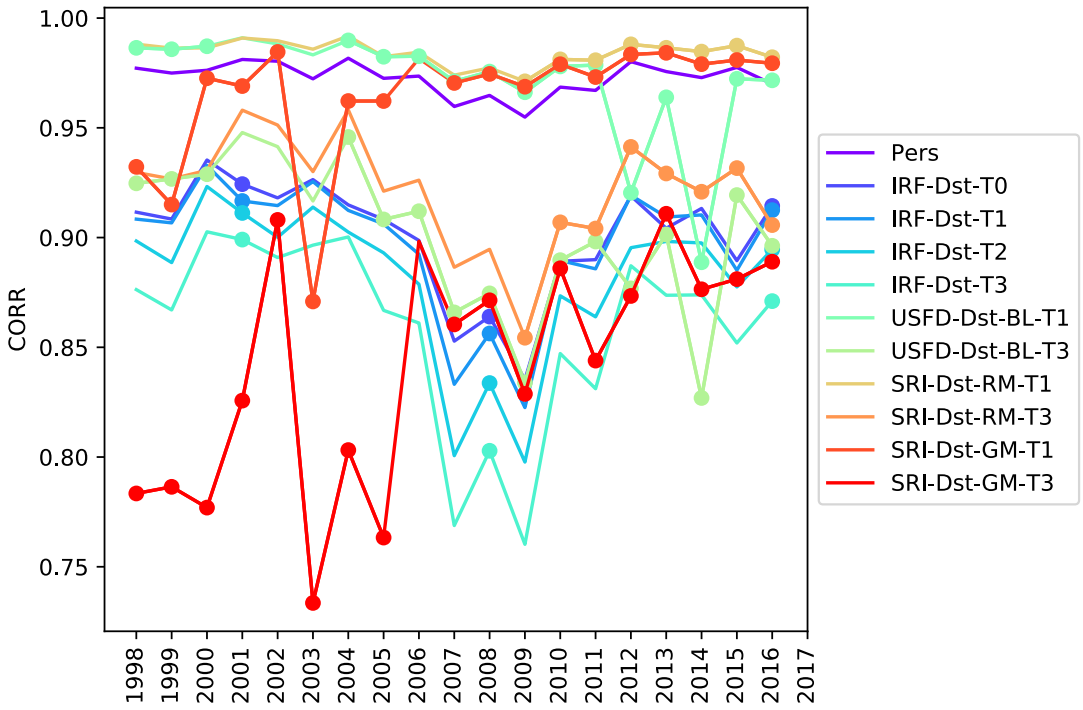


Figure 11: Linear correlation between predicted and observed *Dst* for the different models as function of year. Dots indicate that the year was not part of the training set.

- For *Kp* predictions the IRF-Kp-T0 and T1 models performs best, using past *Kp* seems to have minor effect.
- For *Dst* predictions the SRI-Dst-RM-T1 model performs best, past *Dst* values have significant effect.
- Lead-times beyond 1 hour is generally not possible.
- For real-time implementation, if past indices are used as inputs it will reduce the lead time.

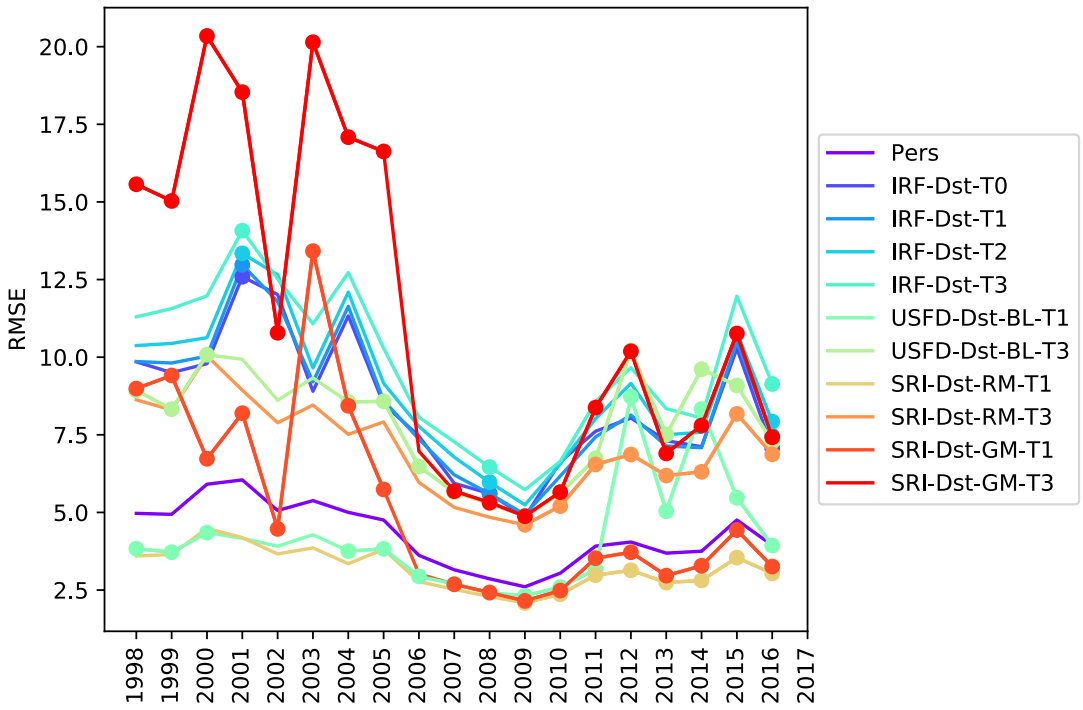


Figure 12: RMSE between predicted and observed *Dst* for the different models as function of year. Dots indicate that the year was not part of the training set.

A GNM models

The result of the GNM models is to generate an upper and lower interval range for the in which the next measurement is expected to lie. These values represent a 90% confidence level.

A.1 Kp 3h ahead

Model inputs	
vBz	solar wind velocity multiplied by IMF Bz component
Kp	previous values of Kp index

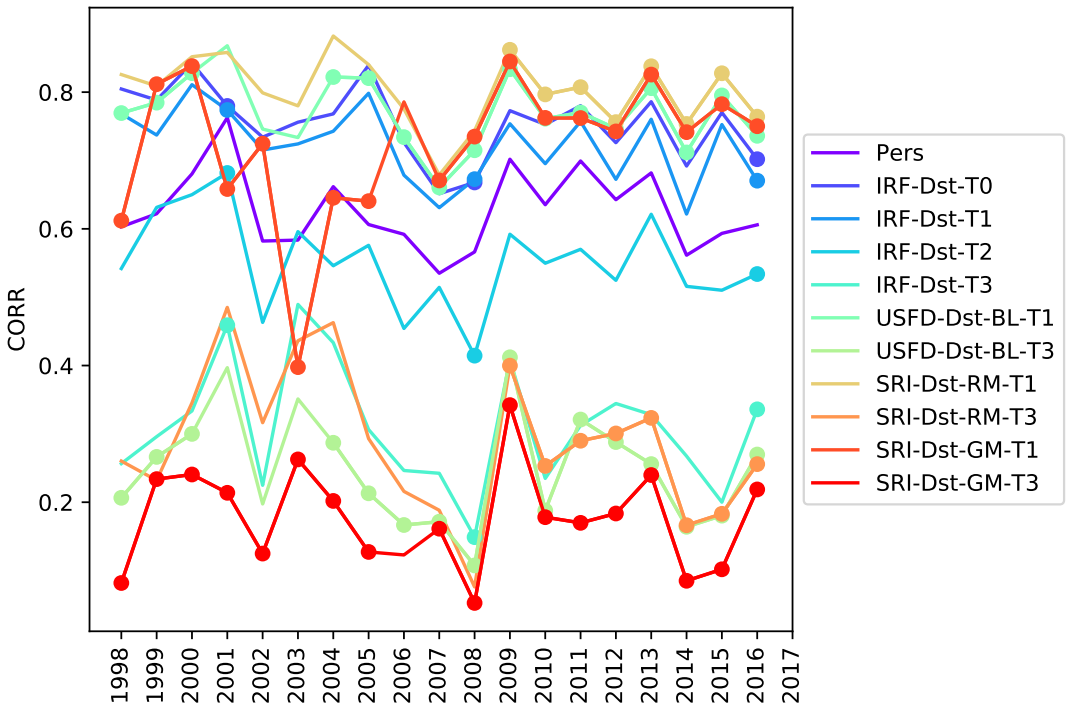


Figure 13: Linear correlation between predicted and observed $dDst/dt$ for the different models as function of year for which $Dst < -13$ nT and $dDst/dt < 0$. Dots indicate that the year was not part of the training set.

$$
\begin{aligned}
Kp_{lower} = {} & 2.0 * (vBz(k) - sigma) + 0.080369 * (kp(k-1)) \\
& - 0.001150 * (kp(k-11)) + 0.663028 * (kp(k)) + 2.857337 * (vBz(k)) \\
& + 0.128944 * (kp(k-2)) + 0.009873 * ((vBz_2(k-2)) * (vBz_1(k-2))) \\
& - 0.017389 * (kp(k-8)) + 0.302823 * (vBz_2(k-11)) \\
& + 0.056715 * ((vBz(k-3)) * (kp(k))) + 0.335488 * (vBz_2(k-12)) \\
& - 0.000269 * ((vBz_2(k-5)) * ((kp(k-2)) * (vBz(k-10)))) \\
& - 0.061579 * (((vBz_2(k-10)) * (vBz_2(k-2))) * ((vBz(k-7)) * ((vBz_2(k-5)) * (vBz_2(k-12)) \\
& + 0.287884 * (vBz_1(k-11)) - 0.068604 * (vBz_2(k-13)) - 0.003930 * (kp(k-13)) \\
& - 2.011700 * (vBz(k-1)) + 0.023151 * ((vBz(k-1)) * (kp(k-2))) \\
& + 0.644566 * (vBz(k-14)) + 0.015417 * ((vBz(k-7)) * (kp(k-4))) \\
& - 0.000781 * ((kp(k)) * (kp(k))) + 0.020964 * ((vBz_1(k-6)) * (kp(k-10))) \\
& + 0.131685 * (vBz(k-10)) - 0.463715 * ((vBz_2(k-4)) * (vBz_2(k-2))) \\
& - 2.549575 * (vBz(k-3)) - 0.523830 * (vBz_1(k-14)) - 0.806390 * (vBz_1(k)) \\
& - 0.624577 * (vBz_1(k-6)) - 0.609556 * (vBz_1(k-1)) - 0.053833 * (kp(k-10)) \\
& + 0.048569 * (kp(k-3)) + 0.179636 * (vBz_2(k-2)) - 0.113916 * (vBz(k-12)) \\
& - 0.023299 * (kp(k-6)) - 0.399621 * (vBz_2(k-9)) - 2.877131
\end{aligned}
$$

$$
\begin{aligned}
Kp_{upper} = {} & 2.0 * (vBz(k) + sigma) + 0.080369 * (kp(k-1)) - 0.001150 * (kp(k-11)) \\
& + 0.663028 * (kp(k)) + 2.857337 * (vBz(k)) + 0.128944 * (kp(k-2)) \\
& + 0.009873 * ((vBz_2(k-2)) * (vBz_1(k-2))) - 0.017389 * (kp(k-8)) \\
& + 0.302823 * (vBz_2(k-11)) + 0.056715 * ((vBz(k-3)) * (kp(k))) \\
& + 0.335488 * (vBz_2(k-12)) - 0.000269 * ((vBz_2(k-5)) * ((kp(k-2)) * (vBz(k-10)))) \\
& - 0.061579 * (((vBz_2(k-10)) * (vBz_2(k-2))) * ((vBz(k-7)) * ((vBz_2(k-5)) * (vBz_2(k-12)) \\
& + 0.287884 * (vBz_1(k-11)) - 0.068604 * (vBz_2(k-13)) - 0.003930 * (kp(k-13)) \\
& - 2.011700 * (vBz(k-1)) + 0.023151 * ((vBz(k-1)) * (kp(k-2))) + 0.644566 * (vBz(k-14)) \\
& + 0.015417 * ((vBz(k-7)) * (kp(k-4))) - 0.000781 * ((kp(k)) * (kp(k))) \\
& + 0.020964 * ((vBz_1(k-6)) * (kp(k-10))) + 0.131685 * (vBz(k-10)) \\
& - 0.463715 * ((vBz_2(k-4)) * (vBz_2(k-2))) - 2.549575 * (vBz(k-3)) \\
& - 0.523830 * (vBz_1(k-14)) - 0.806390 * (vBz_1(k)) - 0.624577 * (vBz_1(k-6)) \\
& - 0.609556 * (vBz_1(k-1)) - 0.053833 * (kp(k-10)) + 0.048569 * (kp(k-3)) \\
& + 0.179636 * (vBz_2(k-2)) - 0.113916 * (vBz(k-12)) - 0.023299 * (kp(k-6)) \\
& - 0.399621 * (vBz_2(k-9)) + 7.1928
\end{aligned}
$$

A.2 Dst 1h ahead

$$
\begin{aligned}
Dst_{lower} = {} & 2 * (vBz(k) - sigma) + 0.005172 * ((index(k)) * (index(k-11))) \\
& + 1.235463 * (index(k) + 0.022015 * (index(k-11)) - 0.341667 * (index(k-1)) \\
& - 0.003082 * ((index(k-11)) * (index(k-2))) + 0.219338 * ((vBz(k-8)) * (vBz(k-2))) \\
& - 1.237752 * (vBz(k)) + 0.145626 * (vBz(k-25)) + 1.452662 * (vBz(k-1)) - \\
& - 0.339149 * ((vBz(k)) * (vBz(k))) + 0.070942 * (vBz(k-24)) \\
& + 0.012110 * ((vBz(k-1)) * (index(k-7))) - 0.001412 * ((index(k-24)) * (index(k-15))) \\
& + 0.044802 * ((vBz(k-25)) * (vBz(k-25))) + 0.109587 * (index(k-14)) \\
& - 0.011974 * ((index(k-11)) * (vBz(k-11))) - 0.025028 * ((vBz(k)) * (index(k))) \\
& + 0.004214 * ((vBz(k-10)) * (index(k-4))) - 0.090366 * (index(k-15)) \\
& + 0.086876 * (index(k-22)) + 0.097186 * (vBz(k-21)) \\
& + 0.002520 * ((index(k-12)) * (vBz(k-12) + 0.012706 * ((index(k-3)) * (vBz(k-1))) \\
& - 0.011400 * (vBz(k-22)) - 0.277522 * (vBz(k-12)) \\
& - 0.003020 * ((vBz(k-22)) * (index(k-17))) + 0.001130 * ((index(k-24)) * (index(k-3))) \\
& - 0.185596 * (vBz(k-5)) + 0.170117 * (vBz(k-8)) + 0.102031 * (vBz(k-23)) \\
& + 0.004543 * ((vBz(k-11)) * (index(k-12))) - 0.051264 * (index(k-7)) \\
& + 0.248782 * (vBz(k-10)) - 0.056410 * (index(k-24)) \\
& + 0.004683 * ((index(k-11)) * (vBz(k-1))) - 0.001967 * ((index(k-7)) * (index(k-11))) \\
& - 0.010614 * ((vBz(k-12)) * (index(k-14))) - 0.204992 * (vBz(k-20)) \\
& + 0.145746 * (vBz(k-3)) + 0.011864 * (index(k-2)) + 0.072516 * (index(k-4)) \\
& + 0.100750 * (vBz(k-19)) + 0.156736 * (vBz(k-14)) - 0.034193 * (index(k-13)) \\
& + 0.173096 * (vBz(k-6)) - 5.0376
\end{aligned}
$$

$$
\begin{aligned}
Dst_{upper} = {} & 2 * (vBz(k) + sigma) + 0.005172 * ((index(k)) * (index(k-11))) \\
& + 1.235463 * (index(k)) + 0.022015 * (index(k-11)) - 0.341667 * (index(k-1)) \\
& - 0.003082 * ((index(k-11)) * (index(k-2))) + 0.219338 * ((vBz(k-8)) * (vBz(k-2))) \\
& - 1.237752 * (vBz(k)) + 0.145626 * (vBz(k-25)) + 1.452662 * (vBz(k-1)) \\
& - 0.339149 * ((vBz(k)) * (vBz(k))) + 0.070942 * (vBz(k-24)) \\
& + 0.012110 * ((vBz(k-1)) * (index(k-7))) - 0.001412 * ((index(k-24)) * (index(k-15))) \\
& + 0.044802 * ((vBz(k-25)) * (vBz(k-25))) + 0.109587 * (index(k-14)) \\
& - 0.011974 * ((index(k-11)) * (vBz(k-11))) - 0.025028 * ((vBz(k)) * (index(k))) \\
& + 0.004214 * ((vBz(k-10)) * (index(k-4))) - 0.090366 * (index(k-15)) \\
& + 0.086876 * (index(k-22)) + 0.097186 * (vBz(k-21)) \\
& + 0.002520 * ((index(k-12)) * (vBz(k-12))) + 0.012706 * ((index(k-3)) * (vBz(k-1))) \\
& - 0.011400 * (vBz(k-22)) - 0.277522 * (vBz(k-12)) \\
& - 0.003020 * ((vBz(k-22)) * (index(k-17))) + 0.001130 * ((index(k-24)) * (index(k-3))) \\
& - 0.185596 * (vBz(k-5)) + 0.170117 * (vBz(k-8)) + 0.102031 * (vBz(k-23)) \\
& + 0.004543 * ((vBz(k-11)) * (index(k-12))) - 0.051264 * (index(k-7)) \\
& + 0.248782 * (vBz(k-10)) - 0.056410 * (index(k-24) \\
& + 0.004683 * ((index(k-11)) * (vBz(k-1))) - 0.001967 * ((index(k-7)) * (index(k-11))) \\
& - 0.010614 * ((vBz(k-12)) * (index(k-14))) - 0.204992 * (vBz(k-20)) \\
& + 0.145746 * (vBz(k-3)) + 0.011864 * (index(k-2)) + 0.072516 * (index(k-4)) \\
& + 0.100750 * (vBz(k-19)) + 0.156736 * (vBz(k-14)) - 0.034193 * (index(k-13)) \\
& + 0.173096 * (vBz(k-6)) + 4.5847
\end{aligned}
$$

B Bilinear models

B.1 Kp 3h ahead

Model inputs

n	density
V	solar wind velocity
$p^{1/2}$	square root solar wind pressure
$BTsin(\theta/2)^6$	

$$
\begin{aligned}
Kp(t) = {} & 3.23556E+00 * BTsin(\theta/2)^6(t-1) \\
& + 5.16727E-01 * Kp(t-3) \\
& + 6.06818E+00 * p^{(1/2)}(t-1) - 2.73663E-04 * V(t-1)Kp(t-3) \\
& - 3.61748E-02 * BTsin(\theta/2)^6(t-1)Kp(t-3) \\
& + 3.40164E-03 * BTsin(\theta/2)^6(t-2)V(t-2) \\
& - 4.72410E-01 * BTsin(\theta/2)^6(t-2)p^{(1/2)}(t-3) \\
& + 1.35797E-04 * V(t-1)Kp(t-6) \\
& - 2.66697E-01 * BTsin(\theta/2)^6(t-6) \\
& - 1.46940E+01 \\
& + 5.95354E-02 * V(t-1) - 2.72031E-02 * V(t-2) \\
& + 1.21319E-01 * BTsin(\theta/2)^6(t-3)p^{(1/2)}(t-2)
\end{aligned}
$$

B.2 Dst 1h ahead

Model inputs

Dst	geomagnetic index Dst
VBs	product of solar wind velocity and southward component of IMF / 1000
p	wind pressure

$$
\begin{aligned}
Dst = {} & [9.5703e-01; Dst(t-1) \\
& - 5.8076e+00 * VBs(t-1) \\
& + 1.7470e+00 * VBs(t-3) \\
& - 1.5876e-01 * p(t-1)VBs(t-1) \\
& - 1.3126e-02 * VBs(t-1)Dst(t-1) \\
& - 1.4909e-01 * VBs(t-1)VBs(t-3) \\
& + 3.2692e+00 \\
& - 7.9361e-02 * p(t-3)VBs(t-3) \\
& + 5.9789e-02 * p(t-2)VBs(t-2) \\
& + 4.8505e+00 * V(t-1)VBs(t-1) \\
& - 6.1593e+00 * V(t-1)
\end{aligned}
$$

B.3 Dst 3h ahead

Model inputs

Dst	geomagnetic index Dst
VBs	product of solar wind velocity and southward component of IMF / 1000
p	wind pressure

$$
\begin{aligned}
Dst = [&8.4798e-01 * Dst(t-3) \\
&- 9.4745e+00 * VBs(t-3) \\
&- 2.3496e+00 * V(t-3)p(t-3) \\
&+ 2.4189e+00 * VBs(t-4) \\
&+ 2.1213e+00 \\
&+ 6.6787e+00 * V(t-6)VBs(t-3)
\end{aligned}
$$

References

Billings, S. A. (2013), *Nonlinear systems identification: NARMAX, Methods in the time, frequency, and spatio-temporal domains*, Wiley-Blackwell.

Billings, S. A., Chen, S. & Korenberg, M. (1989), 'Identification of mimo non-linear systems using a foreward-regression orthogonal estimator', *Int. J. Control* **49**(6), 2157–2189.

Bosworth, J., Foo, N. & Zeigler, B. P. (1972), Comparison of genetic algorithms with conjugate gradient methods, Technical Report CR-2093, NASA.

Chen, S., Billings, S. A. & Lou, W. (1989), 'Orthogonal lease-squares methods and their application to non-linear system identification', *Int. J. Control* **50**(5), 1873–1896.

Eckmann, J. P. & Ruell, D. (1985), 'Ergodic-theory of chaos and strange attractors', *Review of Modern Physics* **57**, 617–656.

Gencay, R. & Dechert, W. D. (1992), 'An algorithm for the n lyapunov exponents of an n-dimensional unknown dynamical system', *Physica D* **59**, 142–157.

Kamide, Y. & Rostoker, G. Eos. 85. No. 19. 188, . (2004), 'What is the physical meaning of the AE index?', *Eos* **85**(11), 188–192.

Koza, J. R. (1992), *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT.

Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J. & Lanza, G. (2003), *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*, Vol. 5 of *Genetic programming*, Springer, US.

Lai, D. & Chen, G. (1998), 'Statistical analysis of lyapunov exponents from time series: A jacobian approach', *Math. Comput. Modeling* **27**(2), 1–9.

Leontaritis, I. & Billings, S. A. (1985*a*), 'Input-output parametric models for nonlinear systems, part i: Deterministic nonlinear systems', *Int. J. Control* **41**, 309–328.

Leontaritis, I. & Billings, S. A. (1985*b*), 'Input-output parametric models for nonlinear systems, part ii: Stochastic nonlinear systems', *Int. J. Control* **41**, 329–344.

Madar, J., Abonyi, J. & Szeifert, F. (2005), 'Genetic programming for the identification of nonlinear inputoutput models', *Ind. Eng. Chem. Res.* **44**(9), 3178–3186.

Mayaud, P. N. (1980), *Derivation, Meaning, and Use of Geomagnetic Indices*, number 22 *in* 'Geophysical Monograph Series', American Geophysical Union, Washington DC.

Parnowski, A. S. (2011), 'Regression modelling of geomagnetic activity', *J. Phys. Studies* **15**(2), 2002.

Rostoker, G. (1972), 'Geomagnetic indices', *Reviews of Geophysics* **10**(4), 935–950.
URL: *http://dx.doi.org/10.1029/RG010i004p00935*

Schweppe, F. C. (1968), 'Recursive state estimation: Unknown but bounded errors and system inputs', *IEEE Trans. Auto. Cont.* **AC-13**(1), 22–28.

Semeniv, O. (2015), 'The combined approach for space weather prediciotn with a guaranteed method and evolutionary algorithm', *J. Phys. Studies* **19**(3), 3003.

Sprott, J. C. (2003), *Chaos and Time Series Analysis*, Oxford University Press.

Temerin, M. & Li, X. (2002), 'A new model for the prediction of *dst* on the basis of the solar wind', *J. Geophys. Res. A* **107**(A12), 1472.

Wei, H. L. & Billings, S. A. (2008), 'Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information', *Int. J. Modelling, Identification and Control* **3**(4), 341–356.

Wei, H. L., Billings, S. A. & Liu, J. (2004), 'Term and variable selection for non-linear system identification', *Int. J. Control* **77**(1), 86–110.