Medium Range Forecasting of Solar-Wind: A Case Study of Building Regression Model with Space Weather Forecast Testbed (SWFT)

Chunming Wang[1]*, I. Gary Rosen[1], Bruce T. Tsurutani[2], Olga P. Verkhoglyadova[2], Xing Meng [2], and Anthony J. Mannucci[2]

[1]Department of Mathematics, University of Southern California Los Angeles, CA 90089, USA.
[2]The Jet Propulsion Laboratory, California Institute of Technology Pasadena, California, USA.

Key Points:

- Space Weather Forecast Testbed (SWFT) jointly developed by USC and JPL provides an useful tool for exploring the possibility of forecasting space weather using data driven techniques.
- Data-driven approaches based on machine learning techniques demonstrate promising performance in solar wind forecast.
- The SWFT is used to streamline the process of developing and validating data-driven forecast models for solar wind speed.

*Sponsorship of the Living With a Star Targeted Research and Technology NASA/NSF Partnership for Collaborative Space Weather Modeling is gratefully acknowledged. Portions of the research for this paper were performed at the Jet Propulsion Laboratory, California Institute of Technology under contract with NASA.

Corresponding author: Chunming Wang, cwang@usc.edu

Abstract

The Space Weather Forecast Testbed (SWFT) is developed by a team of space weather scientists and mathematicians at the University of Southern California (USC) and Jet Propulsion Laboratory (JPL) to foster the creation of models for space weather forecast by exploration of existing historic data using techniques of machine-learning. As an example to demonstrate the potential power of SWFT, we present in this paper a multi-linear regression based forecast model for solar wind. Solar wind is one of the key drivers for numerous physics-based models for space weather including thermosphere and ionosphere models. Many attempts have been made to produce forecasts for the solar wind. SWFT provides an unified framework for forecast model formulation, training and performance assessment. In particular, the preparation of training and validation data by SWFT takes into account of realistic constraints on data latency and forecast lead time. In developing a solar wind forecast model, SWFT allows fast exploration of many meta-parameters such as the list of predictive variables and their time history used in constructing a model. We present the impact of meta-parameter selection, as well as, performance relative to existing solar wind forecast models.

1 Introduction

The wide reliance on wireless communication systems such as the Global Positioning System (GPS), in every aspect of social-economic life and for national security, has highlighted the need for an ability to forecast significant disturbances in the Earth's thermosphere and ionosphere. For the last two decades, significant progress has been made in developing ionosphere data assimilation systems (R. Schunk et al., 2004), (Wang et al., 2004). These systems leverage an abundance of remote sensing data consisting of slant path total electron contents (STEC) to produce an accurate estimation of the current distribution of electrons and ions in ionosphere, i.e., electron density profile (EDP). As in most troposphere numerical weather prediction models, the forecast for future conditions is produced by numerically integrating in time the physics-principle based dynamic model. However, unlike the troposphere, plasma systems in the ionosphere are strongly driven by external driving forces such as solar irradiation fluctuation, perturbation of magnetic and electric fields and, in polar regions, energetic particle precipitation. As a result, the current ionosphere models can only produce reliable short-term forecasts of up to 3 hours lead time. On the other hand, for many practical applications, a medium range forecast with one to three day lead time predicting substantial disruptions in thermosphere and ionosphere is needed. Therefore, while current data assimilation technique represents a milestone in space weather forecast, complementary approaches must be explored to achieve medium range forecast in space weather. At least from a theoretical perspective, the medium range forecast goal could be achieved by stringing together models from the surface of sun, through the interplanetary medium to the Earth's thermosphere and ionosphere. This is indeed what our team proposed to do as a part of NASA and NSF's Living-with-a-star (LWS) program (Mannucci et al., 2015). Although we have achieved some success in identifying sensitivity of thermosphere and ionosphere features to solar-wind in episodic events (Meng et al., 2016), it is also evident that many models for space weather have not been sufficiently calibrated to produce satisfactory predictions. Furthermore, it is also apparent that the lack of longitudinal persistent investigations of ionospheric variability under a wide array of solar and interplanetary conditions makes it difficult for us to precisely characterize ionosphere anomalies in general and those attributable to solar events such as large coronal mass ejections (CMEs) in particular. This realization has prompted us to examine historical data for guidance (Wang et al., 2016). The development of metrics such as n-cluster radius for characterizing the degree to which the state of the ionosphere deviates from nominal condition can be viewed as a form of feature extraction from large dimensional data such as the Global Ionosphere Map (GIM) which is produced at the the Jet Propulsion Laboratory operationally since the early 1990s. Our focus on space weather data and machine-learning techniques led

us to develop the Space Weather Forecast Testbed as a platform to explore the possibility of forecasting space weather with a combination of data and first principle model based approaches. As a demonstration of the utility of SWFT in developing a data based empirical forecast model, we present our experience of forecasting solar wind speed V_x in this paper.

It is well-known that solar events such as coronal-mass-ejections, high-speed solar wind streams and interplanetary magnetic storms can have significant impact on Earth. In particular, the ionosphere conditions critically depend on the variation of solar irradiation, solar wind and variations in the interplanetary magnetic field. A multitude of empirical (Bilitza & Reinisch, 2008),(Y. Chiu, 1975) and first principle based models have been developed (Ridley et al., 2006), (Qian et al., 2014), (R. W. Schunk, 1988),(Huba et al., 2002), (Wang et al., 2004). All these models rely on the availability of key driver parameters such as F10.7, solar wind speed, Kp or Ap index that characterize the space environment. It is common knowledge that the ability to generate medium range forecast for ionosphere conditions critically depends on our ability to forecast the *driver* parameters. In the case of forecasting solar-wind speed, significant research efforts have been reported (Owens et al., 2008), (Owens et al., 2017),(Rotter et al., 2012), (Robbins et al., 2006),(Wintoft et al., 2017),(Vršnak et al., 2007),(Jian et al., 2016),(M. S. Lang et al., 2017), (Henley & Pope, 2017). The earlier efforts were mostly based on a combination of empirical or physics based models (Owens et al., 2008). More recently, data assimilation models have been applied to physics-based space weather models (Henley & Pope, 2017),(Owens et al., 2017),(M. S. Lang et al., 2017). Machine-learning based techniques are also used to classify solar events (Camporeale et al., 2017). Performance of solar-wind forecast models have also been compared (Owens et al., 2008),(Jian et al., 2016). The focus of our study is on the use of tools in SWFT to explore a large set of possible meta-parameters for developing a data-driven model for solar wind forecast. The use of regression analysis technique is one of possible machine-learning techniques that can be helpful in training forecast models.

An outline of the manuscript is as follows. In Section 2, we present the general framework and basic vocabulary for a statistical inference model and their adaptation for a forecast model. We also present the basic components and capability of the SWFT to help its users to explore the space of meta-parameters for developing a forecast model. In Section 3, we present our approach in selecting covariate variables for forecasting V_x. In Section 4, we present comparison of performance of 11 models we constructed. Finally, in Section 5, we present the future directions for SWFT development in general and possible improvement in data driven solar-wind forecast models.

2 Space Weather Forecast Testbed (SWFT)

In this section, we first define a basic statistical framework for machine-learning and its use in the context of developing a forecast model to clarify terminologies that we shall use throughout this manuscript. Then in subsection 2.2, we shall provide an overview of the current state of the SWFT. Finally, in subsection 2.3, we shall present our vision of SWFT as a community tool for space weather forecast research.

2.1 Statistical Framework for Machine-Learning and Forecast

In the general terminology of statistics, an inference model for a quantity Y from the values of variables $X_k, k = 1, \cdots, m$ is a function F that maps X_k to Y. The variable Y is referred to as the dependent variable and the variables $X_k, k = 1, \cdots, m$, are called either covariates, independent, or explanation variables. The process of formulating and calibrating an inference model using dataset

$$D_T = (Y^i, X_1^i, \cdots, X_m^i), i = 1, \cdots, N_T$$

consists of selecting F from a class $\mathcal{F}$ of functions that shows the greatest agreement with available training-data, or equivalently, least discrepancy between Y^i and $F(X_1^i, \cdots, X_m^i)$. This process is generally referred to as regression analysis. The quality of function F with respect to training-data is often measured by

$$\sum_{i=1}^{N_T} L(Y^i - F(X_1^i, \cdots, X_m^i)), \tag{1}$$

where L is sometimes called a *penalty* or *loss function*. In a general statistical framework, the variables Y and $X_k, k = 1, \cdots, m$ are assumed to be random variables following a joint probability distribution $p_{Y,X_1,\cdots,X_m}$ and the training dataset D_T consists of a set of random samples of size N_T where each member $(Y^i, X_1^i, \cdots, X_m^i)$ of the set is sampled independently of other members and follows the distribution $p_{Y,X_1,\cdots,X_m}$.

As a result, the optimal selection F which minimizes (1) is itself a random function dependent on D_T. Moreover, a prediction $F(X_1, \cdots, X_m)$ for a given $X_1, \cdots, X_m$ is also a random variable that depends on the training data set D_T. The ultimate goal of developing a predictive model is to find F that gives value $F(X_1, \cdots, X_m)$ close to Y in a statistical sense. For example, we may wish that the expected value of the prediction error

$$E_{D_T,(Y,X_1,\cdots,X_m)}(Y - F(X_1, \cdots, X_m))$$

is zero. We need to be reminded that in this notation, the expected value is taken over all possible datasets D_T and $(Y, X_1, \cdots, X_m)$.

Definition 1. (Unbiased predictor) The process of obtaining the predictor F from a dataset D_T is unbiased if the above expected value is equal to zero.

In fact, for any specific realization of dataset D_T and therefore a specific realization F of the inference model, we cannot usually establish full statistical properties of $Y - F(X_1, \cdots, X_m)$ without knowing the joint probability distribution $p_{Y,X_1,\cdots,X_m}$. In practice, after the actual value for Y^a is measured, the difference $Y^a - F(X_1^a, \cdots, X_m^a)$ can be determined. This difference should be considered as one realization of the random quantity $Y - F(X_1, \cdots, X_m)$.

Definition 2. (Prediction Error) The difference $Y^a - F(X_1^a, \cdots, X_m^a)$ is called the *prediction error* for predictor F evaluated on data point $(X_1^a, \cdots, X_m^a, Y^a)$.

As in all statistical studies, the distribution of prediction error can be assessed with a new set of data $D_V = (\hat{Y}^j, \hat{X}_1^i, \cdots, \hat{X}_m^j), j = 1, \cdots, N_V$, which is not used for training F, by examining the differences $\hat{Y}^j - F(\hat{X}^j, \cdots, \hat{X}_m^j)$. The dataset D_V is referred to as the validation dataset.

Definition 3. (Machine Learning Algorithm) We call a machine learning algorithm an algorithm that constructs an inference model F using a dataset D_T.

A typical machine learning application involves not only selection of an algorithm for constructing F, but more critically selection of the appropriate forecast variable Y and covariates $X_1, \cdots, X_m$, preparation of the training dataset D_T, and validation dataset D_V. Indeed, for almost all machine-learning applications, these latter tasks are often the most time-consuming ones.

When the above general framework of statistical inference is applied to forecasting the dynamical environment of space weather, a much broader set of questions are often asked. These include:

a. What are physical quantities in space weather that can be reliably predicted and for which their prediction is valuable either for specific applications or for increasing our understanding of underlying physics?

b. How long a lead time for a forecast is feasible and valuable?
c. What are currently measured physical quantities that are the most promising covariates to use in the development of a forecast model for the variables identified in a.?
d. What are impacts of latency of the covariate physical quantities?
e. Should only the latest available measurement of the covariates physical quantities be used in a forecast model or should older measurements also be used?

2.2 Current Status of Space Weather Forecast Testbed (SWFT)

One major goal of the SWFT is to provide an easily accessible platform for users to address the above questions and to explore forecasting strategies for space weather. We are particularly interested in the developments of models that can help to forecast anomalies in the Earth's ionosphere using measurements of solar wind, interplanetary magnetic field (IMF) and other relevant space weather parameters. The three key components of SWFT are

- An extensive database of historical measurements that are quality controlled and registered on a common temporal grid;
- Basic utilities for selecting and assembling data needed for forecast experiments.
- A collection of useful machine learning algorithms and associated analysis tools that can be readily applied to the database of historical measurements.

The current version of SWFT contains 12 years of space weather related data. The data can be divided into three broad groups.

1. Space and Earth based measurements of solar, IMF and geomagnetic activities;
2. Key ionosphere characteristics derived from the Global Ionospheric Map (GIM) which have been produced by the Jet Propulsion Laboratory every 15 minutes for the past 2 decades.
3. A collection of categorical flags derived from data in groups 1 and 2 indicating anomalies and extreme events. In many cases, these flags indicate whether or not a given threshold for the corresponding variable in group 1 or 2 is exceeded. For an example, one of the variables in this group indicates whether or not global maximum VTEC is in the top 5% of historically recorded maximal VTEC value.

The initial selection for the content of the SWFT database aims to provide sufficient and commonly used datasets to enable the experimentation with machine-learning algorithms for space weather forecasting. Our data selection is far from exhaustive and involves many practical trade-offs. In order to allow quick adoption of the most successful and accessible machine-learning algorithms we limit our selection of data to sources where comparable, uniform quality and spatial coverage of data is available over an extensive period of time. This consideration unfortunately makes it impossible to include some of the most common ionosphere measurements such as F2-region peak electron density, NmF2 and its height, HmF2 derived from ionosondes. The inclusion of the third group of variables is motivated by the fact that many machine-learning algorithms are highly effective at producing categorical forecasts (Mitchell, 1997). In fact, it is quite intuitive that in the absence of sufficiently accurate and detailed measurements of relevant space weather parameters, it may only be possible to produce categorical predictions such as most likely all-clear or increased chance of anomaly conditions. A complete list of the dataset in the current version of SWFT is given in Tables 1 and 2.

From its conception, we viewed SWFT as useful for developing forecasting models following the paradigm of *forecast experiment*. In such an experiment, we imagine that we are at a specific time in the past we refer to as the *current epoch*. Our objective is

Table 1. Datasets in current version of SWFT. For data with higher time resolution than 3 hour, statistics of values in each 3 hour time intervals are provided. These are referred as local statistics.

Name	Source	Original Resolution	Local Statistics
Kp	NGDC,NOAA	3 Hrs	
Ap	NGDC,NOAA	3 Hrs	
Cp	NGDC,NOAA	3 Hrs	
SunSpot	SEC, NOAA	3 Hrs	
F107	SEC, NOAA	3 Hrs	
Dst	WDCG, Kyoto	1 Hrs	median,min,max,var
Bx	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
By	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Bz	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Solar-wind speed Vx	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Solar-wind speed Vy	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Solar-wind speed Vz	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Proton density	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Temperature	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Flow pressure	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Ae	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Al	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Au	OmniWeb/GSFC, NASA	5 Min	median,min,max,var
Pcn	OmniWeb/GSFC, NASA	5 Min	median,min,max,var

to train a forecast model for a specific environmental variable of interest with a given lead time, using data available prior to the current epoch. For example, if we want to forecast the solar-wind component V_x one day ahead and we select a current epoch of May 1st, 2011, the possible training data for the forecast model are all data available prior to May 1st, 2011. Suppose that the measurement of V_x is only available one-day after it is measured, that is V_x data has 1-day latency, then the most recent available data is from April 30, 2011. The situation can be further complicated by the availability of covariate data. For example, if we also decide that we need to use the AL index as a covariate and it also has a latency of one days, then the first available training data pair $((V_x, AL))$ consists of V_x measurement for April 30, 2011 and AL index for April 28. This is because that to forecast one-day ahead for the value of V_x in April 30, 2011, the only data we can use are those available on April 29. The latest AL value available on this day is for April 28. A diagram illustrating the chronological relationship between training and validation data for a forecast experiment is shown in Figure 1.

The design of SWFT is to provide maximum flexibility in exploring strategies for space weather forecast. Traditionally an empirical model such as IRI-2012, MSIS or GPS IONO Model for F10.7 (Bilitza & Reinisch, 2007), (Picone et al., 2002),(Klobuchar, 1987) is a parametric model with a limited set of parameters defining the season, geographic location and general space environmental conditions defined by values such as F10.7 and Ap index. These models are typically trained with extensive historical data covering a wide range of conditions spanning multiple solar cycles. A counterpart for the empirical models is a data assimilation model. Typically, at each step, a data assimilation model recursively uses the most recent observation data to optimally initialize a physical law based model. A forecast is produced by propagating the initial state of the model in time. We envision models developed in SWFT can potentially be a hybrid between a paramet-

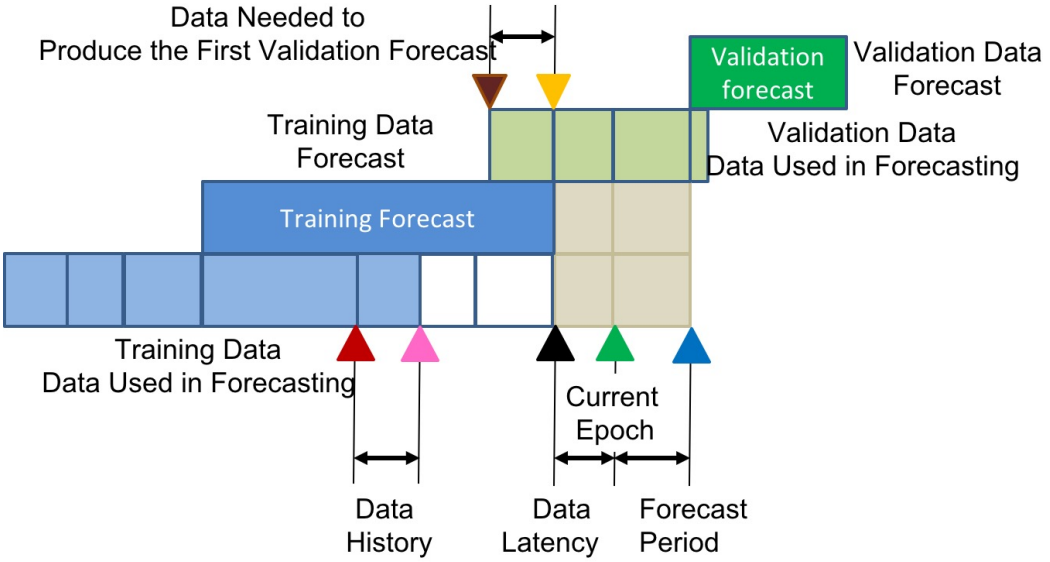

Figure 1. Scope of data usage for a forecast experiment in SWFT. The green triangle at bottom marks the current epoch. The blue triangle on the bottom indicates the first desired forecast value. The time interval between these two triangles corresponds to the forecast lead period. The black triangle on bottom represents the time stamp of the most recently available observational data. The time interval between this data point and current epoch corresponds to data latency. To train a regression model using historical data, the most recent observation for the forecast variable is therefore marked by the third red triangle from right. The total amount of training data for the forecast variable is represented by the middle blue bar. The needed data for the covariate variables must be shifted by the forecast lead time and data latency. The total amount of historical covariate data needed for model training is represented by the bottom blue bar. For an assumed current epoch, the data can be used for validation follows a similar logic. The most recent data for the forecast variable is the same as the first desired forecast. Therefore, the total amount of validation data, represented by the top green bar starts at the same point marked by the right most red triangle on the bottom. The total amount of covariate variables needed to generate the validation values for the forecast variable is represented by the green bar in the middle.

Table 2. Datasets represening ionospheric conditions in the current version of SWFT are mostly derived from the Global Ionospheric Map (GIM) continuously produced by JPL since mid-1990s. The two types of datasets are included. Ionosphere signatures such as global or high-latitude region maximum VTEC or characteristics of equatorial region. The second category consists of cluster radius which measures unusualness of a GIM relative to other GIMs using specific metrics (Wang et al., 2016). Since GIMs are produced every 15 minutes, only the local statistics of data are included in SWFT.

Ionosphere Signatures			
Name	Region	**Signature**	**Local Statistics**
MaxVTECGlobal	Global	Maximal VTEC	median,min,max,var
MaxVTECHiLatNorth	High latitude North	Maximal VTEC	median,min,max,var
MaxVTECHiLatSouth	High latitude South	Maximal VTEC	median,min,max,var
GapML	Equatorial	Mean gap width	median,min,max,var
GapMedian	Equatorial	Median gap width	median,min,max,var
Asym	Equatorial	VTEC asymmetry	median,min,max,var
GIM Cluster Radius			
Name	Region	**Metric**	**Local Statistics**
GIM L_1 HiLat	High latitude	L_1	median,min,max,var
GIM L_1-LocalTime	Local time	L_1	median,min,max,var
GIM L_1-MetricComponents	Global	L_1	median,min,max,var
GIM L_1-Relative	Global	L_1-Relative difference	median,min,max,var
GIM L_1-dLat	Global	L_1-Latitude gradient	median,min,max,var
GIM L_1-dLat LocalTime	Local Time	L_1-Latitude gradient	median,min,max,var
GIM L_∞-HiLat	High latitude	L_∞	median,min,max,var
GIM L_∞-LocalTime	Local time	L_∞	median,min,max,var
GIM L_∞-MetricComponents	Globale	L_∞	median,min,max,var
GIM L_∞-Relative	Globale	L_∞-Relative difference	median,min,max,var
GIM L_∞-dLat	Global	L_∞-Latitude gradient	median,min,max,var
GIM L_∞-dLat LocalTime	Local time	L_∞-Latitude gradient	median,min,max,var

ric empirical model and an assimilation model in the sense that these models rely on recently available data to capture the near-term trend in the relationship between forecast variable and its covariates and, use the detected trend to propagate the state in time.

2.3 SWFT as a Community Tool

The ultimate success of SWFT requires acceptance and participation by the community. SWFT currently consists of openly accessible Matlab source code that reads from a custom multi-year data file formatted as Matlab native binary. The SWFT source code is planned for open access using tools that allow wide dissemination and collaborative development. Matlab statistical and machine learning algorithms are currently the basis for SWFT calculations. It would be relatively straightforward to add additional functionality to SWFT by implementing additional algorithms from the various Matlab toolboxes.

We envision that contributions to SWFT are possible by space weather scientists, as well as, scientists from other disciplines. We feel that such contributions will be greatly

beneficial to both the contributors and the broader community. On the one hand, the contributors of data and algorithms to SWFT will gain wider exposure for their work by allowing a wide cross-section of the community to explore a broad range of uses for their data and techniques. On the other hand, scientists and researchers who are not necessarily knowledgeable in all aspects of space science have the opportunity to explore a large variety of forecast strategies. In fact, the research reported in this paper is an illustration of the potential for using SWFT as a platform for community-wide collaboration: specialized knowledge in measuring and modeling of the solar wind is not required to conduct the experiments, even though solar-wind is one of the key drivers for first principle based ionosphere models such as GITM (Ridley et al., 2006) which we use in our research on ionospheric disturbances.

The data sets that may contribute usefully to SWFT are subject to a set of specific requirements. Such data must be defined on a common time grid with 3 hour resolution. We recognize that inevitable gaps exist in most of dataset and different machine-learning algorithms are designed to handle the issues of missing or imperfect data. A value of NaN is used to indicate missing data in the SWFT database. However, most current algorithms in SWFT do not address general data wrangling. Data wrangling, which refers to quality control of data and methods for recovering missing data, is widely acknowledged as necessary for machine learning algorithms (McGranaghan et al., 2018). Data wrangling can be implemented in future versions of SWFT. However, in general the datasets in SWFT are quality-controlled and have been pre-screened prior to inclusion into SWFT.

Following the general statistical framework for machine-learning, data points in SWFT are assumed to be independent identically distributed samples of an underlying random physical process in some broad sense. As a result, the existing SWFT algorithms should not be directly applied to spatio-temporal data sets such as AMPERE, SuperDARN or satellite data that are acquired at varying locations and times. To bring such data into SWFT requires that the spatio-temporal data are reduced to a time series of physically meaningful quantities. For example, SuperDARN high latitude convection maps have been used to derive the cross polar cap potential (CPCP) or quantity of open magnetic flux at high latitudes (Liu et al., 2019),(Sotirelis et al., 2017). CPCP and magnetic flux are time series that can be used by SWFT. AMPERE field-aligned current maps have been used to derive time series parameters such as hemispheric power and polar cap potential (R. M. Robinson et al., 2018), (R. Robinson et al., 2019). SWFT has applied similar approaches to the spatio-temporal GIM by performing spatial feature extraction (Table 2). Indeed we recognize the challenge in producing time series of space weather data conducive for machine-learning and the development of forecast model. While SWFT only gives a guideline for the form of final dataset to be integrated into its database, the value that the platform SWFT provides to data contributors is that these carefully constructed time series data can be widely used. The sharing of these time series not only greatly reduces redundant research efforts, it also allows detection of possible artifacts due to differences in data preparation methods.

Another challenge for SWFT users is that data quality may vary over time. For example, the reduction of spatio-temporal data to a single time series may depend on the orbit history of a satellite providing the data, or the number of ground stations used to create maps of a geophysical quantity. Users of SWFT must be aware that the results of forecast experiments will depend to some degree on data quality. There are currently no algorithms in SWFT to automatically characterize the degree to which degraded data quality may affect forecasts. With SWFT, it is straightforward to conduct the same experiment over different epochs, e.g. during solar minimum versus solar maximum. If forecast accuracy varies between these two epochs, SWFT cannot distinguish whether this is due intrinsically to sun-Earth connection physics, or to data quality. Users of SWFT must characterize their data quality independently of SWFT and use that information in the design of forecast experiments.

In the Section 3, we shall illustrate the approach of developing a forecast model using SWFT by attempting to forecast V_x with at least one-day lead time.

3 A Solar-wind Forecast Model

Solar wind velocity with its 3 components V_x, V_y and V_z is measured by the Solar Wind Electron Proton Alpha Monitor (SWEPAM) on the ACE spacecraft since 1998 (Mccomas et al., 1998),(M. C. Chiu et al., 1998). Extensive research has shown that solar wind measurements at the L1 Lagrangian point contain key information on the eventual impact of CMEs at Earth. As a result, solar wind velocity is a key input to several thermosphere and ionosphere models (Ridley et al., 2006),(Meng et al., 2016). Numerous studies have shown that solar wind velocity is an important variable to consider when significant ionospheric perturbations occur in the ionosphere. The ability to forecast solar wind velocity is considered a promising step toward developing a medium range forecast for anomalies in the ionosphere.

There have been many research efforts in developing predictive models for solar wind velocity (see (Owens et al., 2008)). Many of these efforts involve physics based modeling of the solar corona and the heliosphere. There are also efforts in making use of the empirical relationship between coronal imagery characteristics and solar wind velocity (Robbins et al., 2006). More recently, data assimilation models combining physics based coronal model with ensemble Kalman filter techniques have been explored (M. S. Lang et al., 2017), (M. Lang & Owens, 2019). Machine-learning techniques have also been used in classification of noteworthy solar-wind anomalies (Camporeale et al., 2017).

In this paper, we develop a statistical regression based solar-wind forecast model using SWFT. A machine-learning approach can complement first principles approaches and establish a useful performance benchmark for forecasting models. The purpose of this benchmark goes beyond the practical need for ranking the relative performance of these models; it can also serve as an indirect measure of our understanding of the physics underlying changes in the solar-wind. Indeed, it is our expectation that improved understanding of physics involved should be confirmed by enhancement in the accuracy of forecast models.

In developing our forecast strategy, we select several key parameters guided by utilitarian reasons. We explore forecasts with a one-day lead time and we assume there exists a one-day latency for the latest covariate measurements. We have arbitrarily selected the current epoch for our experiments to be May 1st, 2011. However, since the model training process is quite efficient, we expect that in practical use, the approach we developed can be used to re-calibrate the model by training it with data from other epochs, including recent data that is most relevant to forecasting in the present epoch. Therefore, the most relevant parameters in the design of our experiments are the number of days $n_{training}$ of data we use to train the regression model and the number of days $n_{validation}$ we use to validate. For all experiments presented in this paper, we took $n_{training} = 365$ and $n_{validation} = 30$. These parameters are often referred to, in the terminology of machine-learning, as parts of meta-parameters for the models. Although the values of these parameters can seem to be selected arbitrarily, they could have profound effects on the results of experiments. One of the key features of SWFT is to allow users to change the values of these meta-parameters easily so that a large number of possible model development options can be explored.

In our case study of solar-wind forecast, we shall focus our attention primarily on another set of key meta-parameters which are the selection of covariates. Traditionally, the selection of covariates used in an empirical forecast model is mostly guided by our understanding of the physics involved that connects the covariates to the forecast variable. Of course physical principles are ultimately developed from observed correlations

between these quantities. In the machine-learning community, the selection of covariates has great similarity to the feature selection process in which a lower dimensional projection of raw data is first identified and used in subsequent learning. Indeed, we consider all available data as possible covariates. However, identification of the most relevant features to the task of forecasting not only reduces the dimensionality of the model training problem, it also has the effect to make the resulting models more robust to noise in data. In fact, randomized feature selection in which features or covariates are randomly picked has also shown to be critical for algorithms such as random regression tree to achieve asymptotic consistency. That is, under the assumption that historical data represent independent and identically distributed samples of the underlying statistical process, when the amount of data used to train a model tends toward infinity, the resulting inference model converges to the "truth". Since the basic assumption for studying statistical consistency is that the training data sample are independent and identically distributed following the same underlying probabilistic distribution as future data, this assumption can be hard to verify in practical situations. In our case study of solar-wind forecast, we use a feature selection approach based on correlations among the variables.

As indicated in the previous section, the number of possible sets of covariates for V_x forecast is extremely large. In our study, we exclude all variables in the SWFT database derived from GIM since our intend is to use forecast of solar-wind to drive ionosphere forecast models. Even excluding the GIM derived variable, there are 61 variables in SWFT. If only the most recent values for these variables are used, there are as many as 2^{61} possible sets of covariates. Moreover, we could also consider recent history of these variables as possible covariates for our inference model, we could consider all 24 possible values measured during the 3 most recent days (eight values per day) where data are available as possible covariates. This would increase of total number of possible sets of covariates to $2^{61\times 24}$. Identifying a promising set of covariates among the large number of possibilities requires an efficient feature selection approach.

The approach we follow in our experiment consists of the following 3 steps.

1. Identify the leading most contemporary pairwise correlated variables in the SWFT database.
2. For each leading variable identified above, perform autocorrelation analysis to determine the length of "memory" in these variables.
3. Perform regression analysis and pruning of the least significant and reliable variables from the list of covariates.

For the first step, we consider each variable in the SWFT database and evaluate the sample cross-correlation coefficient of each variable with V_x. More precisely, we evaluate the quantity c_{i,V_x} defined for variable X_i by

$$c_{i,V_X} = \frac{1}{\sigma_{X_i}\sigma_{V_x}} \sum_{k=1}^{n} (X_i^k - \bar{X}_i^k)(V_x^k - \bar{V}_x^k), \tag{2}$$

where X_i^k and V_x^k are measurements for X_i and V_x at time k and $\sigma_{X_i}, \sigma_{V_x}, \bar{X}_i, \bar{V}_x$ are the n-sample standard deviation and sample mean for X_i^k and V_i^k with $k = 1, \cdots, n$. The cross-correlation coefficients take on values between -1 and 1 and its absolute value represents the strength of correlation between the two variables. Equation (2) is an estimator of the cross-correlation between two random quantities using a size n sample. Therefore, sample cross-correlation is actually a random quantity dependent on the samples X_i^k and V_x^k. When the number of samples n is small, it is not unusual for two sets of randomly selected values to give non-zero sample cross-correlation coefficient. To determine the confidence we have in the estimator c_{i,V_X}, we calculate the p-value for the estimator. The p-value corresponds to the probability that the sample cross-correlation coefficient takes an absolute value larger than what we obtained from our current sample if in reality the cross correlation between the two variables X_i and V_x is equal to zero.

A not very efficient but intuitive way of estimating the p-value consists of constructing random shuffles of samples $X_i^k, V_x^{j(k)}$ where $j(k)$ is randomly selected among the available values. Each shuffled set of data allows us to calculate a new value for c_{i,V_X}. By repeating the shuffling process a large number of times, we can empirically construct a distribution for the value of c_{i,V_X} when the values of X_i are truly uncorrelated to the randomly picked values of V_x. When the variables X_i and V_x are not correlated, the sample cross-correlation coefficient we obtained with the unshuffled original dataset should be similar to the values of c_{i,V_X} obtained from the shuffled datasets. That is, the portion of shuffled datasets that gives values of c_{i,V_X} larger in absolute value than that obtained with the original dataset is not very small. This implies a large p-value. On the other hand, if the two variables are strongly correlated, the sample cross-correlation coefficients derived from the shuffled datasets would be much smaller in absolute value relative to that obtained with original data. This gives a very small p-value. By limiting our evaluation to pairwise cross-correlation between V_x and X_i measured at the same time instance, we greatly reduce the amount of calculation. However, this approach also has significant shortcomings in identifying the most promising covariates for constructing a forecast model for V_x. First, if our objective is to forecast the value of V_x in the future, the future values of covariate variables are also not available at the present time. A more complex issue is that we are interested in finding the most effective set of covariates for the development of the forecast model. Usually, the ensemble of most correlated variables to V_x is not guaranteed to be the most effective set of covariates. As a result, we use this first step as a screening process to find the most promising candidate for the set of covariates.

Using 365 days of data collected prior to the assumed current epoch of May 1st, 2011, we computed pair-wise sample cross-correlation coefficients and their p-values which were determined using Fisher's distribution method instead of random shuffling for all 60 variables in SWFT not derived from GIM. The variables with largest cross-correlation coefficients and their p-values are shown in Figure 2.

Most of these leading variables are expected from our understanding of physics involved. In particular, variables such as V_z, B_x, B_z are measured by instruments on the ACE spacecraft. It is interesting to note that the Dst index is among the most correlated variables to V_x. Since it is well-known Dst represents the disturbance of geomagnetic field which is strongly affected by solar activities, it is expected that Dst is strongly correlated to V_x. However, we may suspect that Dst might not be an effective covariate for a forecast model for V_x because it is a *downstream* covariate, i.e., variation in Dst is a consequence of changes in V_X. Although in the training of forecast models, we strictly follow the basic rule of only using data available prior to the assumed current epoch, we relaxed this rule in the selection of meta-parameter values. In addition to examining the p-values of the leading cross-correlated variables shown in Figure 2, we would like to find out from historical data in SWFT how persistent is the cross-correlation between V_x and other variables. To answer this question, we identified the 14 leading variables most correlated to V_x using each of 12 years of data in the SWFT. Figure 3 shows that at least 8 variables (represented by whiskers with a circle in the middle) have been among the most correlated variables to V_x in every year. These findings give us high confidence these variables should be our first selection as covariates for a forecast model for V_x.

Another important property of a successful set of covariates for a forecast model is that future values are correlated with values in the past. The auto-correlative property of these variables is an indication of the inherent dynamics in the underlying physical system. Analyzing the length of time the covariates are correlated provides us with tentative selection of the time sample beyond the latest available to use as covariates in a forecast model. Indeed for each candidate variables identified by the contemporary cross-

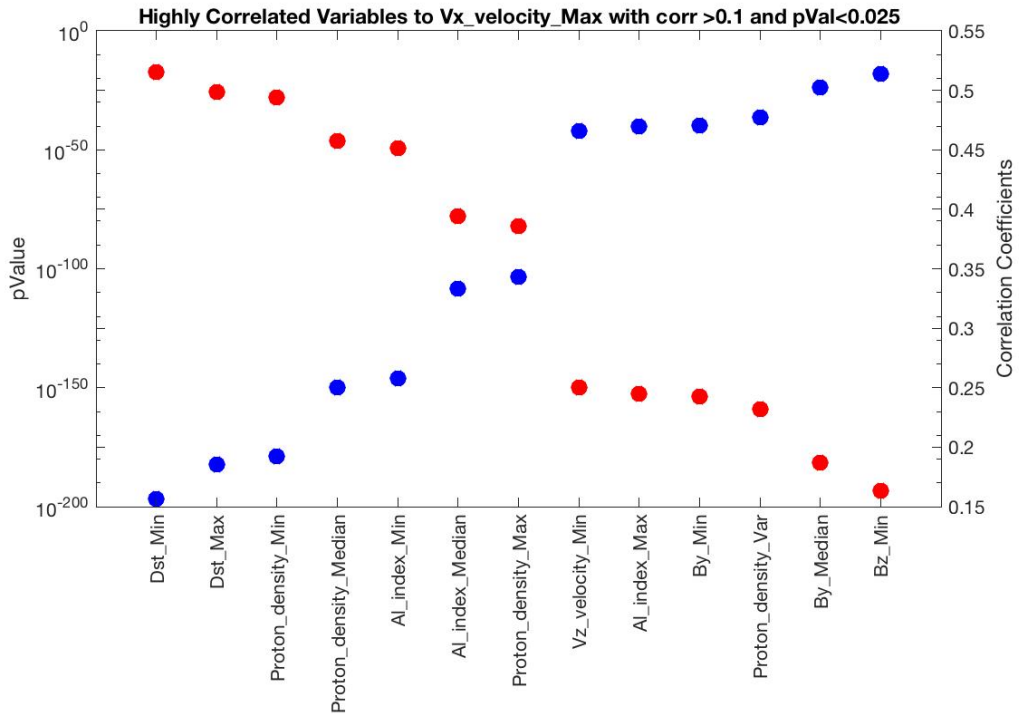

Figure 2. Variables in SWFT with largest cross-correlation coefficient (red-dots) and their p-values (blue-dots) to V_x based on data from April 30, 2010 to May 1st 2011

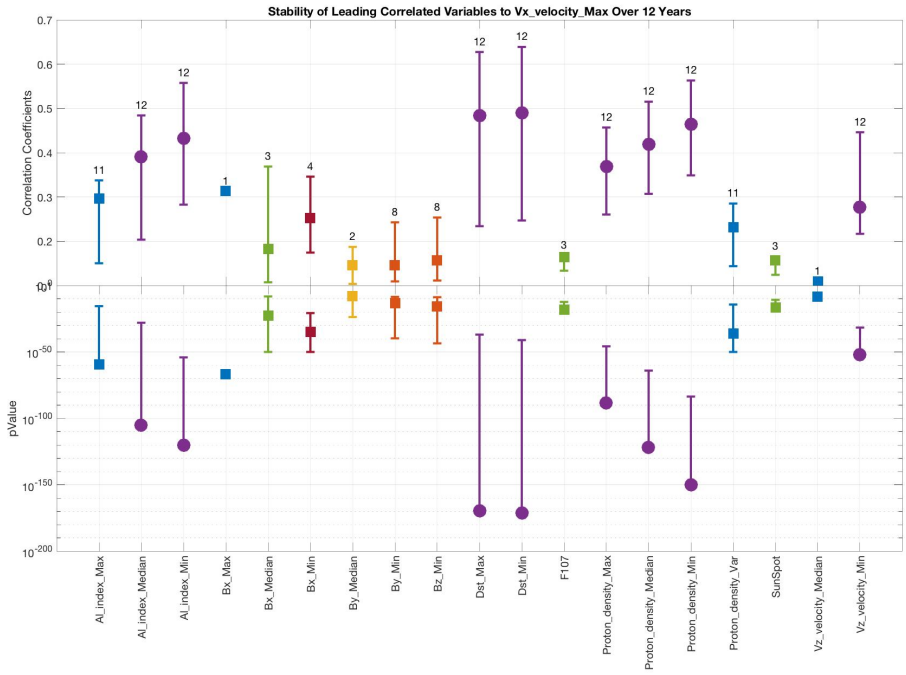

Figure 3. Persistent leading correlated variables in SWFT. The whiskers indicate the range of the cross-correlation coefficients (top) or p-values (bottom) for the variables over 12 years and the squares or circles indicate the median values. The numbers on top of the whiskers and the color indicate the number of years a given variable has been among the most correlated variables.

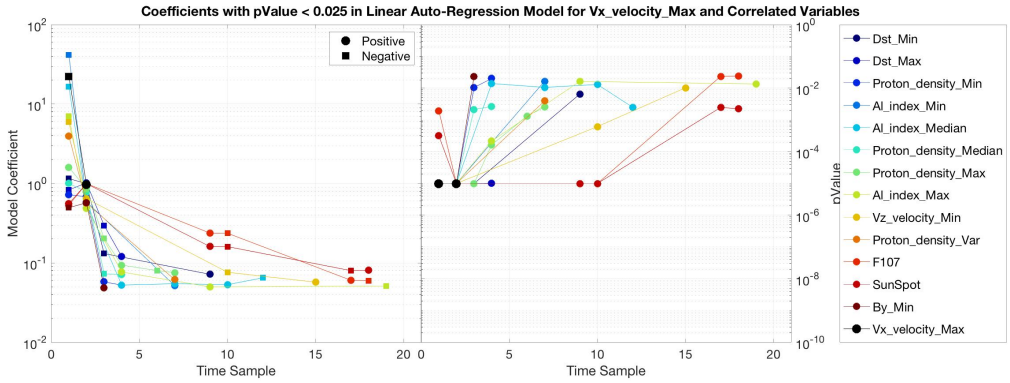

Figure 4. Auto-regression coefficients and their p-values.

correlation analysis, we construct an autoregressive model of the form

$$X_i^k = \sum_{j=1}^{m} a_j X_i^{k-j}. \tag{3}$$

The coefficients $a_1, \cdots, a_m$ are selected so that the following functional is minimized

$$J(a_1, \cdots, a_m) = \sum_{k=1}^{N} |X_i^k - \sum_{j=1}^{m} a_j X_i^{k-j}|^2. \tag{4}$$

Since the optimal values for $a_1, \cdots, a_m$ are functions of sample X_i^k used in the estimation of regression model (3), these values should also be considered as random variables. We can also define the p-values for these variables which represent the probability of obtaining auto-regression coefficients at least as large as we obtain from the training data assuming the data are actually uncorrelated. In Figure 4, the auto-regression coefficients for some of variables identified through contemporary cross-correlation analysis is shown.

It is interesting to note that as the absolute values of the auto-regression coefficients decrease rapidly in time, their p-values also increase. This is consistent with our understanding that most of space weather variables represent aspects of a highly chaotic system. The significant auto-correlations fade rapidly. The results in Figure 4 show that the most meaningful correlation are the among the 16 most recent samples. In all forecast models we only use the latest 2 days of measurements of covariates and V_x.

Once we have identified promising candidate variables in the SWFT database through contemporary cross-correlation analysis and we have established time limits for past values of these covariates to use in a forecast model, we have obtained a tentative list of attributes as input of a prediction model. The data preparation utility in SWFT can be used to extract the appropriate dataset for training and validation. These datasets are collections of pairs of values of V_x at a time k and values of covariate variables X_i at time $k-j$ where j is larger than the sum of forecast lead time and data latency. For example, one of the first forecast models we attempted to construct has properties listed in Table 3.

In this first attempt at developing a forecast model for V_x we used the most recent available values (time stamp -1) and the value for the previous day (time stamp -8) for various local statistics of AL, By, Bz, Dst, proton density and V_z, as well as, the entire available two day history of V_x as covariates. This model has therefore 40 covariates. Since the number of training data points covers 365 days, the values of V_x from May 1st, 2010

Table 3. Attributes of a V_x forecast model Time sample refers to the 3-hour time increment in SWFT. For example, -8 means a 24-hour latent sample (8x3).

Epoch	May 1st, 2011		
Forecast Lead	1 Day	Data Latency	1 Day
Training Data Length	365 Days	Validation Data Length	30 Days

Variable	Covariate ID	Local Statistics	Time Sample
AL	35-36	Max	-1, -8
AL	39-40	Median	-1, -8
AL	37-38	Min	-1, -8
By	5-6	Min	-1, -8
Bz	7-8	Min	-1, -8
Dst	1-2	Max	-1, -8
Dst	3-4	Min	-1, -8
Proton density	27-28	Max	-1, -8
Proton density	31-32	Median	-1,-8
Proton density	29-30	Min	-1,-8
Proton density	31-32	Var	-1, -8
Vx	9-24	Max	-1,...,-16
Vz	25-26	Min	-1, -8

to April 30, 2011 are used as the values for the forecast variable V_x in the training dataset. Since data in SWFT has 3- hour resolution, the training dataset for our first model has 2,920 data points. Associated with each value V_x^k is a vector of length 40 consisting of values of covariates as prepared by utilities in SWFT. The attributes of this vector corresponds to the past values of the covariate variables.

If we denote the components of the covariate vector by X_i, the multi-linear regression model has the form

$$V_x = \sum_{i=1}^{40} \alpha_i X_i + \alpha_0. \tag{5}$$

The multi-linear regression analysis algorithm is used to find coefficients $\alpha_0, \cdots, \alpha_{40}$ such that the following function is minimized:

$$\sum_{k=1}^{N} |V_x^k - \sum_{i=1}^{40} \alpha_i X_i^k - \alpha_0|^2. \tag{6}$$

As we have discussed before, the optimal coefficients $\alpha_0, \cdots, \alpha_{40}$ which are obtained by solving a least squares minimization problem are functions of the random dataset $(V_x^k, X_1^k, \cdots, X_{40}^k)$ for $k = 1, \cdots, N = 2,920$. Therefore, these coefficients should indeed also be considered as random variables. The multi-linear regression analysis algorithm also provides estimation for the p-values for these coefficients. More precisely, Table 4 shows results given by the multi-linear regression algorithm in the Matlab Machine-Learning Toolbox.

The performance of this model is depicted in Figure 5. Figure 5 is one of the standard performance analysis display generated by SWFT. The top panel of the displays shows the comparison between the actual values of V_x (black or blue curves) and the regression model output (magenta or red curves) as a function of time. The region with

Table 4. Results of multi-linear regression analysis. The columns are estimated values for the coefficients (columns 2,7), standard deviation of the estimated value (columns 3,8), t-statistics of the estimated value (columns 4,9) and the p-values of the coefficients (columns 5,10). The coefficient α_0 corresponds to the constant offset in the regression. The coefficient α_i represents the weight for the ith-covariate variable used in this regression where i is the covariate ID given in Table 3

Coeff.	Estimate	SE	tStat	pValue	Coeff.	Estimate	SE	tStat	pValue
α_0	-67.931	9.81	-6.91	5.7e-12	α_{21}	-0.0006	0.10	-0.00	0.99
α_1	-0.9110	0.37	-2.42	0.015	α_{22}	-0.0528	0.10	-0.49	0.62
α_2	-0.0888	0.37	-0.23	0.81	α_{23}	0.1065	0.10	1.00	0.31
α_3	0.6162	0.37	1.65	0.09	α_{24}	-0.0481	0.06	-0.70	0.47
α_4	-0.2310	0.37	-0.61	0.53	α_{25}	0.0738	0.06	1.19	0.23
α_5	2.7755	0.43	6.35	2.3e-10	α_{26}	0.0207	0.06	0.33	0.73
α_6	1.2965	0.44	2.93	0.003	α_{27}	-0.2141	0.63	-0.33	0.73
α_7	5.4892	0.67	8.11	7.2e-16	α_{28}	0.5826	0.66	0.87	0.38
α_8	-0.3405	0.67	-0.50	0.61	α_{29}	-0.7256	1.25	-0.57	0.56
α_9	0.8186	0.07	11.17	2.5e-28	α_{30}	0.7836	1.26	0.61	0.53
α_{10}	-0.1105	0.10	-1.01	0.31	α_{31}	-2.5924	1.16	-2.23	0.02
α_{11}	-0.0034	0.10	-0.03	0.97	α_{32}	0.0280	1.17	0.02	0.98
α_{12}	-0.0132	0.10	-0.12	0.90	α_{33}	0.0153	0.23	0.06	0.94
α_{13}	0.0387	0.10	0.35	0.72	α_{34}	-0.3107	0.24	-1.27	0.20
α_{14}	-0.0710	0.10	-0.65	0.51	α_{35}	0.0149	0.04	0.30	0.76
α_{15}	-0.0207	0.10	-0.18	0.85	α_{36}	0.0534	0.04	1.10	0.27
α_{16}	0.0065	0.10	0.05	0.95	α_{37}	0.0022	0.01	0.16	0.87
α_{17}	0.0561	0.10	0.51	0.60	α_{38}	0.0209	0.01	1.55	0.12
α_{18}	0.0236	0.10	0.21	0.82	α_{39}	0.0081	0.03	0.23	0.81
α_{19}	0.0079	0.10	0.07	0.94	α_{40}	0.0028	0.03	0.08	0.93
α_{20}	-0.0116	0.10	-0.10	0.91					

gray background corresponds to data used to train the regression model. The white background region corresponds to data used for validation which consists of $30 \times 8 = 240$ data points. This plot shows not only that the model is capable of tracking the large variations of V_x over time for the training data, similar quality of prediction can also be achieved for the validation data. The two panels at the bottom reaffirm observations of the top panel. On the left panel, the scatter plots for the true values (horizontal axis) vs. predicted values (vertical axis) of V_x are shown. The blue dots represent the training data and the red dots represent the validation data. As for any scatter plot, the diagonal line corresponds to perfect agreement between true values and model prediction. We can see that the two sets of points (blue and red) have similar distribution with accuracy of prediction degrading for large negative values of V_x compared with performance for lower solar-wind speeds. The advantage of the scatter plots is that they allow us to identify performance of the model for specific ranges of values of the forecast variable. On the other hand, a more summary characterization of model performance can be obtained by the histogram of the prediction errors as shown in the lower right panel. The blue histogram is for the residual of model training. The red histogram shows the prediction error for the validation data. Although the numerical values of mean and standard deviation of prediction error shown in this graph are not unacceptably high, it is interesting to note that the error distribution is skewed toward the negative values.

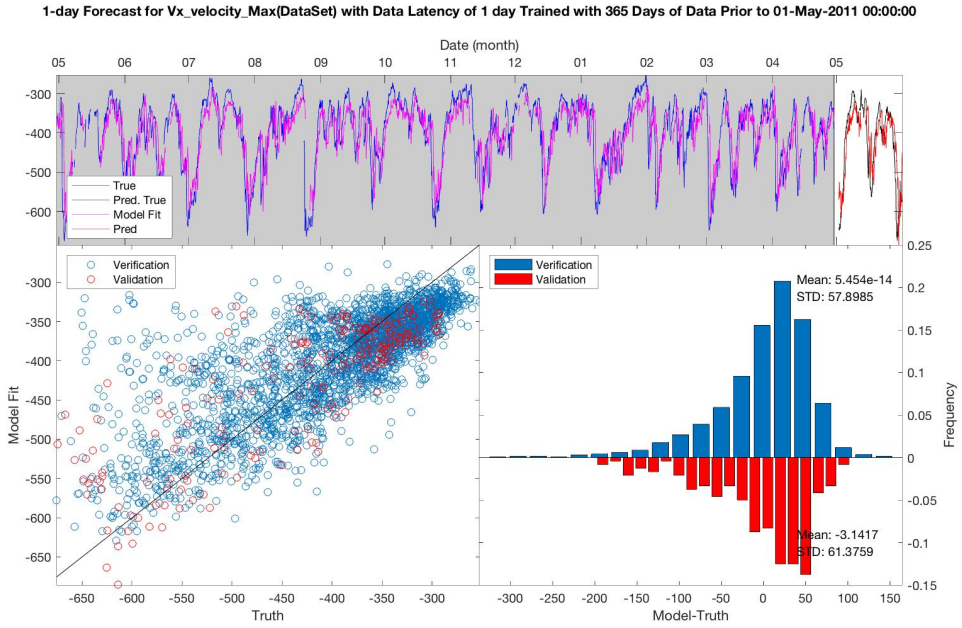

Figure 5. Performance evaluation of the first forecast model. Top panel: Temporal plot of actual data (blue curve on gray background for training data and black curve on white background for validation) and regression model output (magenta curve on gray background for post-fit and red curve on white background for prefit validation). Lower left panel: Scatter plots for training data (blue dots) and validation data (red dots). Lower right panel: Histogram of difference between model output and actual data (blue for training data and red for validation data which is plotted in negative of the frequency)

As a forecast model for V_x with effectively 2-day lead time, the performance for our first model summarized in Figure 5 is not unreasonable. However, we must ask whether all 40 covariates are necessary. An examination of the summary statistics in Table 4 reveals that all but 8 of the coefficients have p-values lower than 0.1. This indicates that the contributions of many covariate variables are not statistically meaningful. In fact, inclusion of unnecessary variables can degrade the performance of a model when applied to validation data because the contribution of coefficients α_i for these variables, while decreasing slightly the post-fit residuals for the training data set, introduces instability to model predictions. This phenomenon is often referred to as "fitting the noise" in an over-fit model training. The last step of constructing an effective forecast model using SWFT involves pruning unnecessary covariates. This effectively involves preparation of new training and validation datasets, running the regression algorithm and evaluating the model performance. Tools in SWFT permit the iteration in model development to be carried out efficiently. In the next section, we shall present our development of a series of models for V_x prediction.

Datasets in SWFT can be broadly separated into physically meaningful measurements or indices and anomaly flags. The latter group of variables are useful if we would like to produce forecasts for nominal vs. anomalous conditions for the space environment or the ionosphere. The forecast models for these anomaly flags are sometimes called classification models. The algorithms for training these models often involve logistic regression or random decision tree/forest (Mitchell, 1997). On the other hand, a wider range of inference models involving nonlinear functions including regression tree-forest type of models are also useful. The underlying data preparation and performance scoring for all these models are similar. The tools in SWFT are crucial for exploration of all these forecast strategy developments.

4 Performance Evaluations

In this section we focus our attention on the detailed selection of meta-parameters for a series of multi-linear regression-based forecast models for V_x. Common features of all these models are: one-day lead time for forecast, one-day data latency, 365 days of training data and 30 days of validation data. Although these parameters may also affect the resulting models, fixing them allows us to better understand the effect of selection of covariates. The 11 models we examine in this section are results of our exploration of different sets of covariate variables in the course of developing a robust forecast model.

One question we would like to answer in the course of our exploration is whether or not our ability to produce reasonable predictions shown in the previous section is primarily due to the inherent dynamical nature of V_x. In other words, since we have used 16 past values of V_x as covariates, we would like to know whether or not introduction of other covariates can substantially improve the model performance. In order to answer this question, we constructed a purely auto-regressive model using past values of v_x only. In Table 4, the coefficients α_9 to α_{24} are associated with past values of V_x. We note that only one of the p-values from this set is lower than 0.1, which is associated with the most recently available value of V_x. As a result, in our new model, we selected only two covariates which are the most recent and one-day earlier values for V_x. The performance summary is shown in Figure 6.

Even though this new model has a drastically smaller number of covariates than the first model presented in the previous section, it is still able to offer comparable predictive accuracy. However, from the top panel, we can already see that the accuracy of model is less satisfactory (with a slightly larger standard deviation for training data and a larger bias for validation data, for example) than for the previous model. This is not surprising because the model has fewer degrees of freedom in fitting the training data. The summary statistics in Table 5 shows that even though in the model presented in the

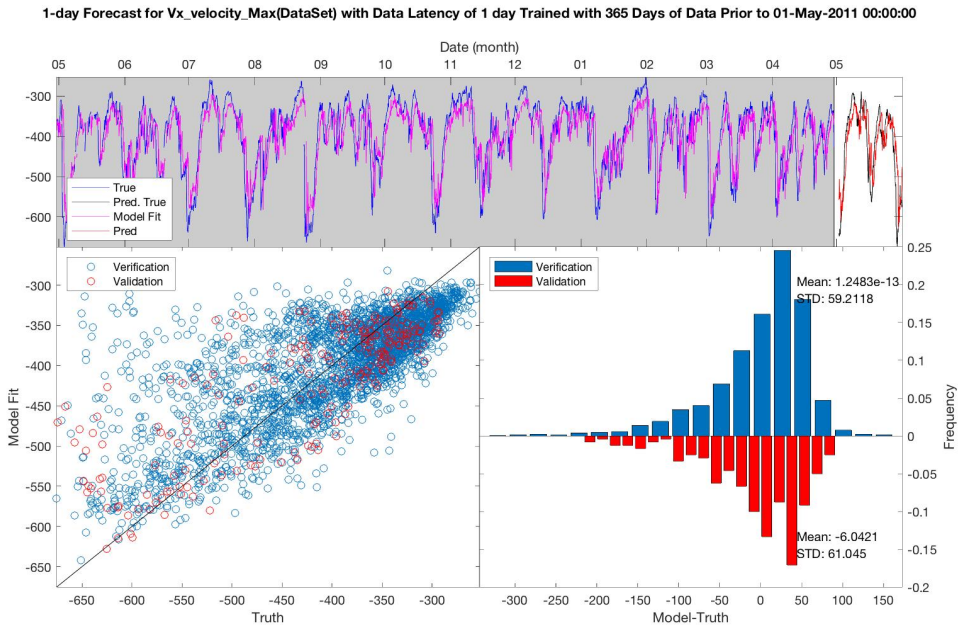

Figure 6. Performance evaluation of forecast model using only past values of V_x as covariates.

previous section, only the most recent values of V_x has significant contribution, in the new model, not only the weight α_2 associated with the one-day old value of V_x in the regression model is not substantially smaller in absolute value than α_1, its p-value is also very small. It is possible that information about future value for V_x in the older values of V_x can also be gained from other covariates in the previous model. Therefore, removal of those covariates from the new model makes the older time sample of V_x more important.

Table 5. Results of multi-linear regression analysis for model using only past 2 values of V_x as covariates. The columns are estimated values for the coefficients (column 2), standard deviation of the estimated value (column 3), t-statistics of the estimated value (column 4) and the p-values of the coefficients (column 5). The coefficient α_0 corresponds to the constant offset in the regression. The coefficient α_i represents the weight for the ith-covariate variable used in this regression. The covariates are the two most recent observation of V_x.

Coeff.	Estimate	SE	tStat	pValue
α_0	-122.78	5.485	-22.384	3.52e-102
α_1	1.2887	0.068586	18.79	2.9898e-74
α_2	-0.59488	0.068599	-8.6718	7.1015e-18

We would like to see how well the value of V_x can be predicted if none of the recent values of V_x are used as covariates. The summary statistics are given in Table 6 and the performance summary is given Figure 7. From 7 we see a substantial reduction in the performance of the model without using any past value of V_x despite of the fact that the p-values of the coefficients are quite small.

Table 6. Results of multi-linear regression analysis for model without past values of V_x as covariate. The columns are estimated values for the coefficients (columns 2,7), standard deviation of the estimated value (columns 3,8), t-statistics of the estimated value (columns 4,9) and the p-values of the coefficients (columns 5,10). The coefficient α_0 corresponds to the constant offset in the regression. The coefficient α_i represents the weight for the ith-covariate variable used in this regression. The covariates in this model consists of all variables listed in Table 3 except past values of V_x and local variation of proton density.

Coeff.	Estimate	SE	tStat	pValue	Coeff.	Estimate	SE	tStat	pValue
α_0	-336.01	4.42	-75.9	0	α_{12}	0.86	0.58	1.4	0.13
α_1	-1.79	0.44	-4.0	5.9e-05	α_{13}	3.03	1.31	2.3	0.02
α_2	-0.64	0.44	-1.4	0.1495	α_{14}	2.97	1.33	2.2	0.02
α_3	2.11	0.44	4.7	1.9e-06	α_{15}	-3.19	1.34	-2.3	0.01
α_4	0.47	0.44	1.0	0.28	α_{16}	-0.46	1.39	-0.3	0.73
α_5	5.25	0.51	10.1	5.8e-24	α_{17}	-0.26	0.05	-4.3	1.2e-05
α_6	3.79	0.51	7.3	3.5e-13	α_{18}	-0.10	0.05	-1.7	0.08
α_7	1.76	0.76	2.2	0.02	α_{19}	0.09	0.01	5.9	2.9e-09
α_8	-0.70	0.77	-0.9	0.36	α_{20}	0.08	0.01	5.1	3.1e-07
α_9	0.52	0.07	7.2	7.6e-13	α_{21}	0.08	0.04	2.0	0.03
α_{10}	0.18	0.07	2.4	0.01	α_{22}	0.04	0.04	1.1	0.26
α_{11}	1.08	0.55	1.9	0.05					

We explored a total of 11 models with different sets of covariates see Figure 8. Each row in the plot on represents one of the 11 models. The color bars in the right panel indicate the variables and their time history used as covariates for the specific model. The left panel provides information on the postfit (red, also referred to as training error which corresponds to the difference between model output and actual training data) and prediction (blue, also referred to as validation error which represents the difference between model output and actual data when model is applied to validation data that is not used to train the regression models) root-mean square error of V_x in km/s. For example, row number 9 corresponds to the first forecast model presented in the previous section and row number 11 corresponds to the 2-term auto-regression model discussed earlier in this section.

Figure 8 shows that the difference between the worst performing model with RMSE of nearly 78 to the best model with RMSE of 58 is over 20 percent despite of the fact that all these models have covariates selected among the most promising variables in SWFT. This suggests the usefulness of exhaustively exploring the meta-parameter space for developing forecast models. More close examination of the relative performance indicates that models with the largest number of covariates such as model 9 has the best fit to the training data, its performance on validation data suggests that other models such as models 6 and 8 with smaller number of covariates can achieve superior performance by eliminating statistically insignificant covariates. In Table 7 we evaluated the Akaike information criterion (AIC) by assuming the residual error between actual data and model predict for both training and validation data follow normal distribution. Since AIC is a relative measure of information uncertainty in model prediction with low value indicating higher quality of prediction (Aho et al., 2014), we can see that model 6 has the lowest AIC value judging from the validation data while model 9 has the lowest AIC value judging from the training data. While it is possible to speculate a reason for the predictive performance of each these regression models, we do not feel that space physics principles favorite any one of these over the others.

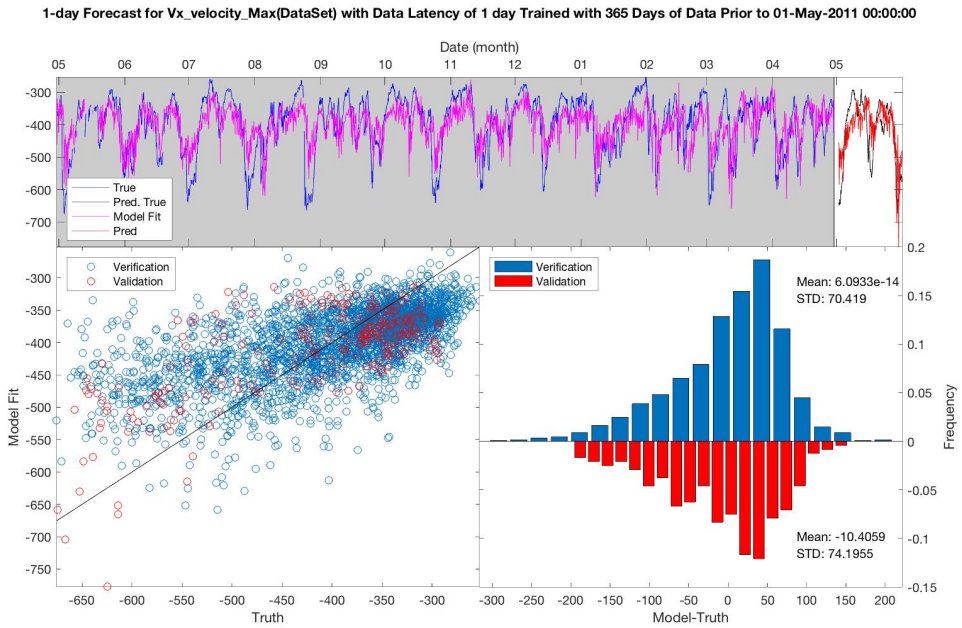

Figure 7. Performance evaluation of forecast model without using any past values of V_x as covariate.

Table 7. Akaike information criterion (AIC) Values evaluated using training and validation data.

Model Number	Training	Validation pValue
1	31856	2715
2	31794	2724
3	30640	2626
4	31662	2752
5	31430	2759
6	27958	2499
7	27946	2513
8	30324	2581
9	27925	2528
10	31185	2717
11	27929	2550

We understand that the comparison between solar wind forecast models presented above with physics based models such as WSA, WSA-ENLIL, and CORHEL is problematic because the physics based models use much smaller set of space environmental parameters to initiate their prediction (Owens et al., 2008). In particular, the amount of data used to train models reported in (Owens et al., 2008) is much more extensive than used by our regression models. The aim of physics based model is also much broader than simply predicting the value of V_x. In particular, by using dataset covering a significant portion of a solar cycle, a physics based model may attempt to represent the inherent variation in behavior of solar wind over an extended period of time. On the other hand, regression based models focus primarily on relatively short-term trends that may be re-

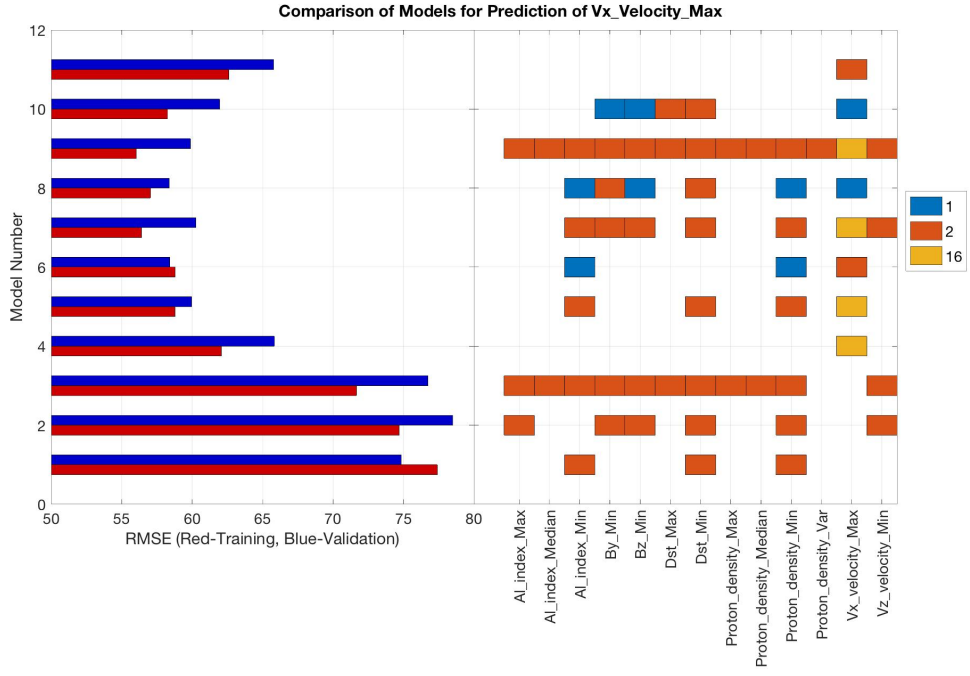

Figure 8. Selection of covariates of 11 forecast models for V_x. Left panel shows the RMSE where red bars represent training or postfit error and blue bars represent validation or prefit error for the 11 models. The right panel indicates the covariates used in each regression model. The colored bars indicate the variables used as covariate for a model and the legends 1, 2, 16 indicate the number of past historical samples used for the variables. That, 1 for when only the most recently available sample is used; 2 for the most two recently available samples were used and, 16 for when all 16 most recently available samples or entire two recently available days of measurements are used.

liable for over specific portion of a solar cycle. Naturally, these models are expected to be retrained on a regular basis to catch the latest trends. However, compared to the RMSE of over 90 km/s (Owens et al., 2008), the data based V_x forecast models are shown to deliver comparable performance as physics based models for solar wind forecast that can be used as inputs for thermosphere and ionosphere models at least for the majority "nominal" conditions.

5 Conclusion

In this manuscript we have presented a case study of constructing a regression based forecast model for solar wind velocity or more precisely V_x using the Space Weather Forecast Testbed. A similar approach can also be used to forecast other components of the solar wind. As illustrated in Section 4, SWFT makes experimentation with a large number of models straightforward. In fact, even though the multilinear regression models presented in this manuscript are commonly used, the same data preparation and performance evaluation tools can be used for a wider range of modeling approaches. In particular, an immediate extension of the models presented here is a regression-tree model. By sub-dividing training data according to the range of the predicted V_x values into smaller subsets, a new multilinear regression model for each subset can be derived. Mathematically, the post-fit residual for the regression-tree model can be reduced to arbitrarily low levels by continuing the sub-dividing process. Of course this does not guarantee that the validation error would be reduced as well. We shall present our the results of these new approaches in a subsequent paper.

It is our belief that the combination of easy access to extensive historic space weather data, data preparation and analysis tools and a wide range of machine-learning algorithms made available by SWFT can play a significant positive role in promoting the development of new space weather forecast models. It is important to note that these models are not limited to purely empirical data-driven models. Computer simulation results using physics based model can just as easily be introduced into the SWFT database and used in constructing forecast models. A new stand-alone version of SWFT will soon be available to the space weather community. It is our wish that more people in space weather community would have the opportunity to experiment with SWFT and help to further develop SWFT. In fact, as a part of community resource, SWFT provides a platform for incorporating much larger variety of space weather observational data than in the current prototype. Inclusion of more extensive data sets such as AMPERE, SuperDARN, Van Allen Probes would necessarily presents new technical challenges which may require expansion of the current framework for SWFT. However, as a starting point, we are convinced that SWFT showed great promise of data-driven forecasting approach using machine-learning techniques.

Acknowledgments

Sponsorship of the Living With a Star Targeted Research and Technology NASA/NSF Partnership for Collaborative Space Weather Modeling is gratefully acknowledged. Portions of the research for this paper were performed at the Jet Propulsion Laboratory, California Institute of Technology under contract with NASA. The authors would also like to express our appreciation for the two reviewers of this manuscript whose suggestions have been extremely helpful for improving the final paper.

The sources of original data used in SWFT are indicated in the manuscript. We are currently working with the Coomunity Coordinated Modeling Center (CCMC) to make the processed data in SWFT database available through CCMC.

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of aic and bic. *Ecology, 95 (3)*. (DOI10.1890/13-1452.1)

Bilitza, D., & Reinisch, B. (2007). International reference ionosphere 2007: Improvements and new parameters. *Advances in Space Research, 42*.

Bilitza, D., & Reinisch, B. (2008). International reference ionosphere 2007: Improvements and new parameters. *Advances in Space Research, 42*. (DOI10.1016/j.asr.2007.07.048)

Camporeale, E., Car, A., & Borovsky, J. E. (2017). Classification of solar wind with machine learning. *journal of Geophysical Research: Space Physics, 122, 10,910–10,920*. (DOI10.1002/2017JA024383)

Chiu, M. C., Von-Mehlem, U. I., Willey, C. E., Betenbaugh, T. M., Maynard, J. J., Krein, J. A., ... Rodberg, E. H. (1998). Ace spacecraft. *Space Science Reviews, 86: 257-284*.

Chiu, Y. (1975). An improved phenomenological model of ionospheric density. *journal of Atmospheric and Terrestrial Physics, 37(12):1563-1570*. (DOI10.1016/0021-9169(75)90035-5)

Henley, E. M., & Pope, E. C. D. (2017). Cost-loss analysis of ensemble solar-wind forecasting: Space weather use of terrestrial weather tools. *Space Weather, 15, 1562–1566*. (DOI10.1002/2017SW001758)

Huba, J., Dymond, K., Joyce, G., Budzien, S., Thonnard, S., Fedder, J., & McCoy, R. (2002). Comparison of o+ density from argos lorass data analysis and sami2 model results. *Geophys. Res. Lett. 29*. (DOI10.1029/2001GL013089)

Jian, L. K., MacNeice, P. J., Mays, M. L., Taktakishvili, A., Odstrcil, D., B. Jackson, H. S. Y., ... Sokolov, I. V. (2016). Validation for global solar wind prediction using ulysses comparison: Multiple coronal and heliospheric models installed at the community coordinated modeling center. *Space Weather, 14, 592–611*. (DOI10.1002/2016SW001435)

Klobuchar, J. (1987). Ionospheric time-delay algorithm for single-frequency gps users. *IEEE Trans. Aerospace and Electronic Sys., AES-23, 325-331*.

Lang, M., & Owens, M. J. (2019). A variational approach to data assimilation in the solar wind. *Space Weather, 17, 59?83*. (DOI10.1029/2018SW001857)

Lang, M. S., Browne, P. A., van Leeuwen, P. J., & Owens, M. (2017). Data assimilation in the solar wind: Challenges and first results. *Space Weather, 15, 1490–1510*. (DOI10.1002/2017SW001681)

Liu, E., Hu, H., Liu, J., Teng, X., & Qiao, L. (2019). Predicting superdarn cross polar cap potential by applying regression analysis and machine learning. *Journal of Atmospheric and Solar-Terrestrial Physics, 193, 105057*.

Mannucci, A. J., Verkhoglyadova, O., Tsurutani, B. T., Meng, X., Pi, X., Wang, C., ... Hajra, R. (2015). Medium-range thermosphere-ionosphere storm forecasts. *Space Weather, 13, 125-129*. (DOI10.1002/2014SW001125)

Mccomas, D. J., Bame, S. J., Barker, P., Feldman, W. C., Phillips, J. L., & Riley, P. (1998). Solar wind electron proton alpha monitor (swepam) for the advanced composition explorer. *Space Science Reviews 86: 563-612*.

McGranaghan, R. M., Mannucci, A. J., Wilson, B., Mattmann, C. A., & Chadwick, R. (2018). New capabilities for prediction of high-latitude ionospheric scintillation: A novel approach with machine learning. *Space Weather, 16*.

Meng, X., Mannucci, A. J., Verkhoglyadova, O. P., & Tsurutani, B. T. (2016). On forecasting ionospheric total electron content responses to high-speed solar wind streams. *J. Space Weather Space Clim., 6(A19)*. (DOI10.1051/swsc/2016014)

Mitchell, T. M. (1997). *Machine learning*. Mcgraw-Hill International Editions Computer Science Series, 1st.

Owens, M. J., Riley, P., & Horbury, T. S. (2017). Probabilistic solar wind and geomagnetic forecasting using an analogue ensemble or "similar day" approach.

Solar Phys. (2017) 292:69. (DOI10.1007/s11207-017-1090-7)

Owens, M. J., Spence, H. E., McGregor, S., Hughes, W. J., Quinn, J. M., Arge, C. N., ... Odstrcil, D. (2008). Metrics for solar wind prediction models: Comparison of empirical, hybrid, and physics-based schemes with 8 years of l1 observations. *Space Weather, Vol. 6, S08001*. (DOI10.1029/2007SW000380)

Picone, J. M., Hedin, A. E., Drob, D. P., & Aikin, A. C. (2002). Nrlmsise-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *journal OF GEOPHYSICAL RESEARCH, VOL. 107, NO. A12, 1468*. (DOI10.1029/2002JA009430)

Qian, L., Burns, A. G., Emery, B. A., Foster, B., Lu, G., Maute, A., ... Wangm, W. (2014). *The ncar tie-gcm: A community model of the coupled thermosphere/ionosphere system, in modeling the ionosphere-thermosphere system.* AGU Geophysical Monograph Series.

Ridley, A. J., Deng, Y., & Toth, G. (2006). The global ionosphere-thermosphere model. *J. Atmos. Solar-Terr. Phys., 68, 839-864*. (DOI10.1016/j.jastp.2006.01.008)

Robbins, S., Henney, C. J., & Harvey, J. W. (2006). Solar wind forecasting with coronal holes. *SolarPhys. 233:265–276*. (DOI10.1007/s11207-006-0064-y)

Robinson, R., Anderson, B., & Zanetti, L. (2019). Ampere-derived electrodynamic parameters of the high latitude ionosphere (adelphi). *Geophysical Research Abstracts, 21, EGU2019-3577*.

Robinson, R. M., Zhang, Y., Anderson, B. J., Zanetti, L. J., Korth, H., & Fitzmaurice, A. (2018). Statistical relations between field-aligned currents and precipitating electron energy flux. *Geophysical Research Letters, 45(17), 8738?8745,*. (doi:10.1029/2018GL078718)

Rotter, T., Veronig, A., Temmer, M., & Vršnak, B. (2012). Relation between coronal hole areas on the sun and the solar wind parameters at 1 au. *Solar Phys., 281:793–813*. (DOI10.1007/s11207-012-0101-y)

Schunk, R., Scherliess, L., Sojka, J., Thompson, D., Anderson, D., Codrescu, M., ... Howe, B. (2004). Global assimilation of ionospheric measurements (gaim). *Radio Sci., 39, RS1S02*. (DOI10.1029/2002RS002794)

Schunk, R. W. (1988). *A mathematical model of the middle and high latitude ionosphere.* PAGEOPH, 1988, 127:255. (DOI10.1007/BF00879813)

Sotirelis, T., Keller, M. R., Liou, K., Smith, D., Barnes, R. J., Talaat, E., & Baker, J. B. H. (2017). Testing the expanding-contracting polar cap paradigm. *JGR, 24(A7), 1905?10*. (doi:10.1002/2017JA024238)

Vršnak, B., Temmer, M., & Veronig, A. (2007). Coronal holes and solar wind high-speed streams: I. forecasting the solar wind parameters. *Solar Phys. 240: 315–330*. (DOI10.1007/s11207-007-0285-8)

Wang, C., Hajj, G. A., Pi, X., Rosen, I. G., & Wilson, . D. (2004). Development of the global assimilative ionospheric model. *Radio Sci., 39, RS1S06*. (DOI10.1029/2002RS002854)

Wang, C., Rosen, I. G., Tsurutani, B. T., Verkhoglyadova, O. P., Meng, X., & Mannucci, A. J. (2016). Statistical characterization of ionosphere anomalies and their relationship to space weather events. *Space Weather Space Clim., 6, A5*.

Wintoft, P., Wik, M., Matzka, J., & Shprits, Y. (2017). Forecasting kp from solar wind data: input parameter study using 3-hour averages and 3-hour range values. *J. Space Weather Space Clim., 7, A29*. (DOI10.1051/swsc/2017027)