Space Weather

RESEARCH ARTICLE
10.1029/2023SW003514

MagNet—A Data-Science Competition to Predict Disturbance Storm-Time Index (*Dst*) From Solar Wind Data

Manoj Nair[1,2], Rob Redmon[2], Li-Yin Young[1,2], Arnaud Chulliat[1,2], Belinda Trotta[3], Christine Chung[4], Greg Lipstein[4], and Isaac Slavitt[4]

[1]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA, [2]NOAA's National Centers for Environmental Information, Boulder, CO, USA, [3]Bureau of Meteorology, Melbourne, VIC, Australia, [4]DrivenData Inc., Denver, CO, USA

Key Points:

- The MagNet challenge attracted 622 participants from 64 countries who submitted 1,197 models to predict Dst in a real-time modeling environment
- The ensemble average of the top four winning models, which used different modeling architectures, performed better than individual models for both competition and post-competition data
- The challenge revealed notable successes and areas for improvement

Correspondence to:
M. Nair,
manoj.c.nair@noaa.gov

Citation:
Nair, M., Redmon, R., Young, L.-Y., Chulliat, A., Trotta, B., Chung, C., et al. (2023). MagNet—A data-science competition to predict disturbance storm-time index (*Dst*) from solar wind data. *Space Weather*, *21*, e2023SW003514. https://doi.org/10.1029/2023SW003514

Received 30 MAR 2023
Accepted 30 AUG 2023

Author Contributions:

Conceptualization: Manoj Nair, Rob Redmon
Data curation: Li-Yin Young, Belinda Trotta, Christine Chung
Formal analysis: Belinda Trotta
Funding acquisition: Manoj Nair
Investigation: Manoj Nair
Methodology: Manoj Nair, Rob Redmon, Belinda Trotta, Christine Chung
Project Administration: Manoj Nair
Software: Li-Yin Young, Belinda Trotta, Christine Chung, Greg Lipstein, Isaac Slavitt
Supervision: Manoj Nair
Validation: Li-Yin Young, Belinda Trotta, Christine Chung, Greg Lipstein, Isaac Slavitt
Visualization: Manoj Nair, Greg Lipstein
Writing – original draft: Manoj Nair, Arnaud Chulliat, Belinda Trotta

Abstract Enhanced interaction between solar-wind and Earth's magnetosphere can cause space weather and geomagnetic storms that have the potential to damage critical technologies, such as magnetic navigation, radio communications, and power grids. The severity of a geomagnetic storm is measured using the disturbance-storm-time (*Dst*) index. The *Dst* index is calculated by averaging the horizontal component of the magnetic field observed at four near-equatorial observatories and is used to drive geomagnetic disturbance models. As a key specification of the magnetospheric dynamics, the *Dst* index is used to drive geomagnetic disturbance models such as the High Definition Geomagnetic Model—Real Time. Since 1975, forecasting models have been proposed to forecast Dst solely from solar wind observations at the Lagrangian-1 position. However, while the recent Machine-Learning (ML) models generally perform better than other approaches, many are unsuitable for operational use. Recent exponential growth in data-science research and the democratization of ML tools have opened up the possibility of crowd-sourcing specific problem-solving tasks with clear constraints and evaluation metrics. To this end, National Oceanic and Atmospheric Administration (NOAA)'s National Centers for Environmental Information and the University of Colorado's Cooperative Institute for Research in Environmental Sciences conducted an open data-science challenge called "MagNet: Model the Geomagnetic Field." The challenge attracted 622 participants, resulting in 1,197 model submissions that used various ML approaches. The top models that met the evaluation criteria are operationally viable and retrainable and suitable for NOAA's operational needs. The paper summarizes the competition results and lessons learned.

Plain Language Summary Solar wind interacting with Earth's magnetosphere can cause geomagnetic storms, damaging critical technologies. The disturbance-storm-time (*Dst*) index measures storm severity, driving geomagnetic disturbance models. Traditional *Dst* forecasting relied on solar wind observations, but recent Machine Learning (ML) models show promise. However, many are unsuitable for operational use. To explore viable ML solutions, National Oceanic and Atmospheric Administration and the University of Colorado organized the "MagNet: Model the Geomagnetic Field" challenge. Six hundred and twenty-two participants submitted 1,197 ML-based models for predicting Dst. The top-performing models meeting evaluation criteria are operationally viable and retrainable, meeting NOAA's operational needs. The paper summarizes competition results and insights, emphasizing ML's potential to enhance geomagnetic storm forecasting for practical applications.

1. Introduction

The coronal mass ejections and corotating interaction regions are considered the main sources of geomagnetic storm (Mursula et al., 2022). During these events, the enhanced solar wind efficiently couples with the Earth's magnetic field. The resulting ground magnetic field variations increase the errors of systems that use the Earth's natural magnetic field as a pointing reference. The Dst or disturbance-storm-time index is a measure of the severity of the geomagnetic storm (e.g., Gonzalez et al., 1990). Geomagnetic storms can disrupt satellite communications, Global Positioning System navigation systems, compass-based pointing systems, and power grids, leading to widespread blackouts and disruptions in communication. As a key specification of the magnetospheric dynamics, the Dst index is used to drive geomagnetic disturbance models such as National Oceanic and Atmospheric Administration (NOAA)/NCEI's High Definition Geomagnetic Model—Real Time (HDGM-RT). The HDGM-RT is a global, high-resolution model of the Earth's geomagnetic main, crustal, and external field, providing magnetic

Writing – review & editing: Manoj Nair, Rob Redmon, Arnaud Chulliat, Belinda Trotta, Christine Chung, Greg Lipstein

field values (total field, dip, and declination) at or near the Earth's surface. The HDGM (Nair et al., 2021) is updated annually to correctly model secular changes in the geomagnetic field. The HDGM-RT is developed by the US NOAA, in partnership with the directional drilling industry and the University of Colorado. The external field of the HDGM-RT is a sum of ionospheric and magnetospheric components derived from an annually updated diurnal magnetic variation model (Chulliat et al., 2013, 2016) and the POMME model (Maus et al., 2010; Maus & Lühr, 2005). Additionally, magnetic surveyors, government agencies, academic institutions, satellite operators, and power grid operators use the Dst index to analyze the strength and duration of geomagnetic storms.

The Dst is calculated as an average value of the horizontal component (H) of the magnetic field observed at four near-equatorial observatories (Hermanus, Kakioka, Honolulu, and San Juan), which are far from the auroral and equatorial electrojets and are located roughly evenly distributed in longitude (Sugiura, 1963). The Dst is calculated using the following steps. First, a baseline is determined for H at each observatory using a quadratic polynomial model based on the annual means of the previous 6 years and then subtracted from H at each observatory. The solar quiet daily variation (Sq) is estimated for each observatory by fitting the Fourier series (for local time and month). The disturbance variation at each observatory is calculated by further subtracting the Sq from the baseline adjusted H. Finally, the Dst index is obtained by averaging the disturbance variation of four observatories, taking into account their latitude in geomagnetic dipole coordinates at every Universal-Time (UT) hour. The range of the Dst index is generally about −400 to +100 nT (nanoTeslas), with large negative values indicating a geomagnetic storm. Extremely intense storms can cause the Dst index to drop below −500 nT. A typical geomagnetic storm can last a few days and can be generally described in three phases. The initial phase, if present, can have short periods (1–3 hr) of positive Dst values due to the sudden compression of magnetosphere. The main phase of a geomagnetic storm can last from a few hours to as long as 24 hr or more, when Dst reaches its lowest values. The recovery phase is characterized by a gradual increase in the Dst index toward its normal or baseline level and typically lasts from several hours to a few days. The observed Dst is the sum of the magnetic fields from external (Est, ring current) and internal (Ist, induced counterpart) sources. For driving models such as HDGM-RT, the Est is separated from Dst by using a method described by Maus and Weidelt (2004). The World Data Center (WDC) Kyoto (Japan) provides the official Dst index (Sugiura, 1963) at three processing levels: Final, Provisional, and Quicklook. Currently, the Final version is available through the year 2016 and is their most-processed version, using definitive magnetic field data. The Provisional version is available for the years 2017–2019 and uses preliminary magnetic field data. Quicklook is their near real-time version. All of these processing levels are available with a 1-hr time resolution. For real-time operations purposes, the Quicklook Dst is often used.

Empirical models have been proposed as early as 1975 (Burton et al., 1975) to forecast Dst solely from solar wind observations at the Lagrangian (L1) position. Several models were proposed for solar wind forecasting of Dst. They include empirical (e.g., Bala & Reiff, 2012; Burton et al., 1975; Lundstedt et al., 2002; O'Brien & McPherron, 2000; Temerin & Li, 2002) and physics-based (e.g., Raeder et al., 2001; Tóth et al., 2005) models. The Geospace Environment Modeling challenge of 2008–2009 asked modelers to submit Dst results for four geomagnetic storm events and five types of observations that can be modeled by empirical, climatological, or physics-based models of the magnetosphere-ionosphere system (Rastätter et al., 2013). The results showed that during the peak of geomagnetic storms, empirical models (including a neural network-based model) performed better than physics-based models. However, empirical models struggled to accurately predict the "quiet-time" Dst baselines. Over the past decade, various machine-learning (ML) based Dst specification models have been proposed (Chandorkar et al., 2017; Cristoforetti et al., 2022; Gruet et al., 2018; Laperre et al., 2020; Lazzús et al., 2017; Tasistro-Hart et al., 2021) as an improvement over traditional Dst models. All of these examples except Lazzús et al. used solar wind data to forecast Dst. And most of these past efforts relied on prior Dst timesteps for the forecast step, which is generally not available in real-time.

Despite the large amount of research in *Dst* forecasting, it is difficult to compare different prediction methods against each other. Researchers implement their prediction methods using a subset of data, using specific computational platforms and optimized for certain periods of solar cycles. While most of these efforts aim to accurately predict the *Dst* values, their use in real-time operational environments is limited by three factors. (a) Most models use the prior *Dst* values as input (in addition to the solar wind parameters) to predict the future *Dst* values. However, the aforementioned Quicklook *Dst* is derived from unverified raw data, and it often contains inaccurate values caused by spikes, noise, and baseline shifts. In general, the Quicklook *Dst* is available with a latency of about 2 hr. However, there is no guaranteed latency for the Quicklook *Dst*, and it can be delayed by several hours. In Figure 1, we plot the Quicklook Dst (real-time) and the provisional *Dst* (released about a year later) for 1 June 2020, through 10 August 2020. For the entire period of July 2020, the Quicklook *Dst* had a mean bias error of about

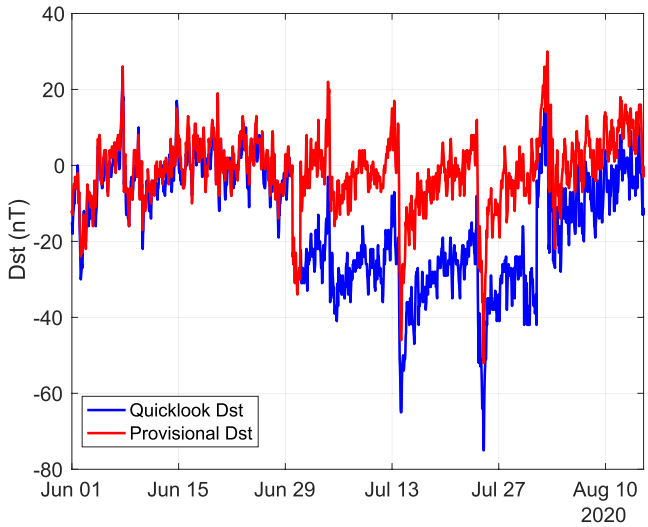

Figure 1. A comparison of the real-time, "Quicklook" *Dst* (less accurate) and provisional *Dst* (more accurate). The provisional *Dst*, released typically with 1-year latency does not contain the baseline errors seen in the Quicklook *Dst*.

−25 nT. Hence, the operational dependency of a model on prior *Dst* values is detrimental to their forecasting reliability. (b) Most of the models are developed using solar wind data available from the NASA OMNI website (e.g., Papitashvili & King, 2020). These data sets are often "time-shifted" to account for travel time to the magnetosphere's bow shock nose (from their measurement location at L1) and have undergone processing to remove noise and other non-geophysical contamination (NASA, 2022). Additionally, the OMNI data set may contain data from non-operational satellites such as WIND. In contrast, the Real-Time Solar Wind (RTSW) data from NASA Advanced Composition Explorer (ACE) and NOAA DSCOVR satellites are provided without being time-shifted and may contain noise, data gaps, and spikes owing to sensor and processing system malfunctions. Ideally, a real-time *Dst* predictor should internalize the time shift/propagation delay. (c) The models should be able to be run in an operational environment with limited, operationally approved computational resources and using the operational "Real-Time Solar Wind" (RTSW, Zwickl et al., 1998) data stream. Thus, for operational use, it is important to develop a model that (a) does not depend on the past *Dst* values, (b) is agnostic to the noise and data gaps in the RTSW data, and (c) can run on a specific computational environment using RTSW data sets.

1.1. Why a Data Science Competition?

We attempted to solve this problem by conducting an open competition among data scientists. We are motivated by this question: Can data scientists with or without prior training in geophysics improve *Dst* forecasting? We were optimistic for three reasons. First is that competition platforms such as Kaggle, TopCoder, and DrivenData have demonstrated how data science can be successfully outsourced to people without domain expertise. Many organizations have run competitions on such diverse topics as right whale identification (Bogucki et al., 2019), optimizing flight routes, predicting ocean health, and diabetic detection (Graham, 2015). To our knowledge, this is the inaugural instance of a data science competition being conducted to address a problem in the field of space physics. Data scientists with little or no expertise in the domain have responded brilliantly with useful solutions. Second is that the democratization of data-science tools and the availability of cloud-based resources for modeling enabled more people to take part in such competitions. For example, machine-learning frameworks such as TensorFlow and PyTorch are open-source and publicly available. Serious machine-learning model development is possible using free online notebooks such as Google's Colaboratory—a browser-based front-end editor with a cloud-based backend for data processing. Finally, data-science platforms attract scientists and engineers with prize money and the prestige of winning the competition. More importantly, their leaderboards have become central to job placement in the data science industry (Martinez & Walton, 2014). Another advantage of soliciting modeling solutions from a wider pool of solvers (as against using in-house developers) is that the former has the potential to bring in a diverse set of strategies to achieve the same goal. Specifically, it is known that using an ensemble of diverse models is often better than relying on individual models (e.g., Boukabara et al., 2020; Riley et al., 2013; Weyn et al., 2021). An open data science competition has the potential to provide several high-quality models but using different modeling strategies. The NOAA National Centers for Environmental Information (NCEI), in partnership with the University of Colorado's Cooperative Institute for Research in Environmental Sciences (CIRES) and the NASA Center of Excellence for Collaborative Innovation (CoECI), conducted an open data science challenge "MagNet" to forecast *Dst* using the solar wind data from 15 December 2020, through 12 February 2021, aligned with NOAA's new Artificial Intelligence Strategic goals to advance AI research and strengthen and expand partnerships (NOAA, 2021). The competition was implemented by DrivenData and HeroX. We describe the data sets provided to the competitors in Section 4, the competition progress in Sections 5 and 6, and the evaluation of the top models in Section 7. Using a post-competition RTSW data set (March 2021 through May 2022), we assess the performance of the top models in Section 5, well after the competition was completed with newer geophysical data.

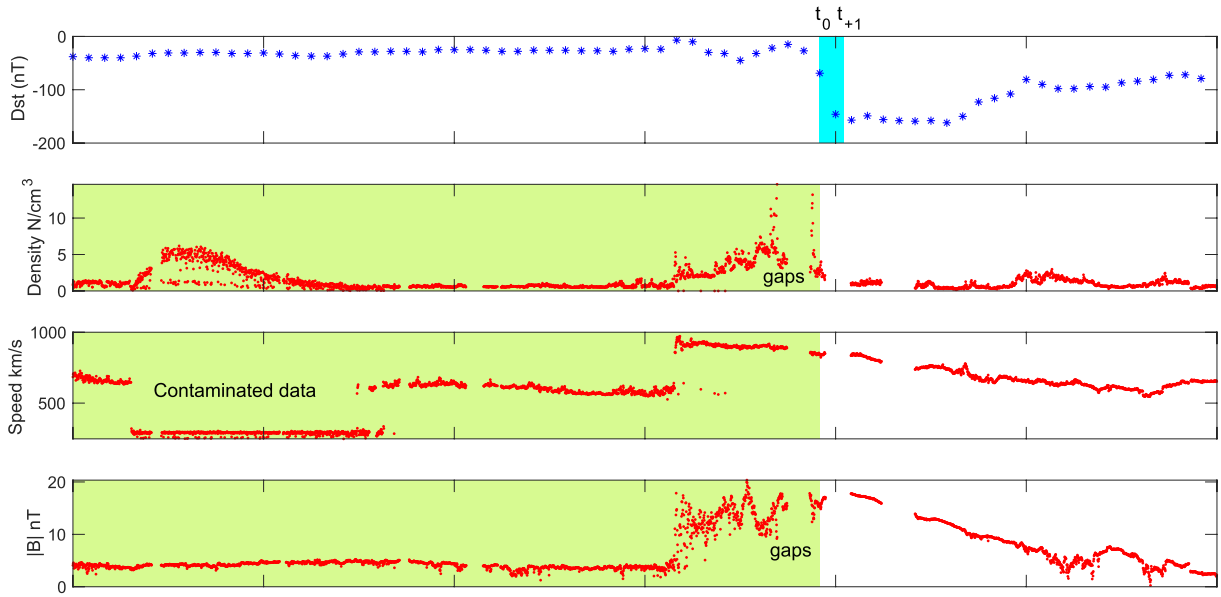

Figure 2. Given 1-min averages of Real-Time Solar Wind (RTSW) measurements in the window t_{-N} to t_0, where N is $7 \times 1{,}440$ corresponding to 1 week of 1-min averages, predict *Dst*, at t_0 and t_{+1}. The historical solar wind data are shaded in green and the *Dst* prediction is given in blue. The RTSW data used for the competition may contain gaps and errors.

2. Pre-Competition Phase

The proposal for the "MagNet" challenge was competitively selected for funding through the 2020 *NCEI Innovates program* The program is designed to encourage exploration and cross-center interaction that could result in outcomes that help NCEI achieve its missions. Since NOAA/NCEI does not have a mechanism for conducting the competition, an Interagency Agreement (IAA) was signed with NASA. In fulfillment of the IAA, the NASA Tournament Lab selected DrivenData as the competition platform provider through an open process. Regular meetings were held between DrivenData, NCEI, and NASA in preparation for the competition to finalize the data sets, performance metrics, competition rules, and judgment criteria. Before the competition set dates, the outreach team of HeroX reached out to the data-scientist community. The NCEI/CIRES team communicated the opportunity with the space-physics community via emails and personal contacts. The concept of the challenge was presented and discussed at the Workshop on ML, Data Mining, and Data Assimilation in Geospace (LMAG2020) in the fall of 2020. The authors of recent papers on *Dst* forecasting were directly contacted to encourage their participation.

3. Problem Statement

Figure 2 shows an overview of the problem statement. Given the 1-min averages of RTSW solar wind data for the past 1 week, along with the optional satellite position and sunspot data predict the *Dst* values at present and 1 hr in the future (t_0 and t_{+1}). Note that the RTSW data may contain gaps and noise, so the solvers will need to come up with strategies to deal with them.

4. Data Set for the Competition

The following section describes the "official" data set provided to the competition.

4.1. Solar-Wind Data

The input data for the modeling challenge are the solar wind data measured by NASA's ACE (launched in 1997) and NOAA's Deep Space Climate Observatory (DSCOVR, launched in 2015), situated at approximately 1.6 million kilometers from the Earth and orbiting the Lagrangian (L1) position. The L1 point is a neutral gravity

Table 1
Solar Wind Data Available for Use in the Modeling Challenge

Variable ID	Variable	Description	Units	Min/Max
1	time_delta	Time delta from the start of a segment, for example, 27 days 08:00:00		–
2	bx_gse	Interplanetary magnetic field (IMF) X-component in Geocentric solar ecliptic (GSE) coordinates	nT	−200/+200
3	by_gse	Interplanetary magnetic field Y-component in GSE coordinates	nT	−200/+200
4	bz_gse	Interplanetary magnetic field Z-component in GSE coordinates	nT	−200/+200
5	theta_gse	Interplanetary magnetic field latitude in GSE coordinates (defined as the angle between the magnetic vector B and the ecliptic plane, being positive when B points North)	Degrees	−90/90
6	phi_gse	Interplanetary magnetic field longitude in GSE coordinates (the angle between the projection of the IMF vector on the ecliptic and the Earth-Sun direction)	Degrees	0/360
7	bx_gsm	Interplanetary magnetic field X-component in Geocentric solar magnetospheric (GSM) coordinates	nT	−200/+200
8	by_gsm	Interplanetary magnetic field Y-component in GSM coordinates	nT	−200/+200
9	bz_gsm	Interplanetary magnetic field Z-component in GSM coordinates	nT	−200/+200
10	theta_gsm	Interplanetary magnetic field latitude in GSM coordinates	Degrees	−90/90
11	phi_gsm	Interplanetary magnetic field longitude in GSM coordinates	Degrees	0/360
12	bt	Interplanetary magnetic field magnitude	nT	0/200
13	MAG Source	Starting in 2016, the solar wind data at any timestamp can be sourced from either DSCOVR or ACE satellites depending on availability and quality. Plasma and MAG source vectors are the same.	1 = ACE, 2 = DSCOVR	1/2
14	Density	Solar wind proton density	N/cm^3	0/200
15	Speed	Solar wind bulk speed	km/s	200/2,000
16	Temperature	Solar wind ion temperature	Degrees K	1.00E4/1.00E7
17	Plasma Source	Starting in 2016, the solar wind data at any timestamps can be sourced from either DSCOVR or ACE satellites depending on the quality	1 = ACE, 2 = DSCOVR	1/2

point between the Sun and the Earth that is about a hundredth of the distance to the Sun. L1 is a good position from which to monitor the Sun because the constant stream of particles from the Sun (the solar wind) reaches L1 about an hour before reaching the Earth. The solar wind data consists of in situ measurements of magnetic field and plasma. Specifically, we use the "Real-Time Solar Wind" (RTSW) product of NOAA's Space Weather Prediction Center (SWPC), in 1-min averages, for the years 1998 through 2020. The RTSW data for the past 7 days are available, in real-time, from the website https://services.swpc.noaa.gov/products/solar-wind/. The RTSW data older than 7 days are available on request from SWPC. The RTSW data include the interplanetary magnetic field (IMF), solar wind speed, density, and temperature measurements transmitted from the L1 position in near real-time. The RTSW product is delivered, in real-time by SWPC and is used for space-weather operational models such as the Coupled Thermosphere Ionosphere Plasmasphere Electrodynamics Model (CTIPe, Codrescu et al., 2012) and the Prompt-Penetration Electric Field Model (PPEFM, Manoj & Maus, 2012). Unlike the solar wind data provided by NASA's OMNI website–commonly used for developing *Dst* forecasting models–the RTSW data is not time-shifted from L1 to the bow-shock nose of the magnetosphere. During times of outage or other problems in DSCOVR data, SWPC uses data from the NASA/ACE spacecraft. Hence, at any point in time (starting in 2016), the RTSW may source data from ACE or DSCOVR and they are indicated as such. Table 1 provides a summary of the solar wind data set used in this modeling challenge.

4.2. Spacecraft Position Data

The ACE and DSCOVR satellites are not stationary at the L1 point. They orbit around the L1 point, in a relatively constant position to the Earth as the Earth revolves around the Sun. The positional information might give additional improvements to the forecasting of the *Dst* values. The daily positional information of the satellites in Geocentric solar ecliptic Coordinates are described in Table 2.

Table 2
Spacecraft Location Available for Use in the Modeling Challenge

Variable ID	Variable	Description	Unit	Min/Max
18	time_delta	Time delta from the start of a segment, for example, 27 days 08:00:00	–	16 February 1998 00:00:00.000/10 November 2020 05:29:00.000
19	GSE_X (km)	Position of the satellite in the X direction of GSE coordinates	Kilometers	85,516/1,594,772
20	GSE_Y (km)	Position of the satellite in the Y direction of GSE coordinates	Kilometers	−475,678/267,959
21	GSE_Z (km)	Position of the satellite in the Z direction of GSE coordinates	Kilometers	−161,542/164,061

4.3. Sunspot Numbers

The Sun exhibits a well-known, periodic variation of the number of spots on its disk over a period of about 11 years. The solar wind data set provided to the competition spans Solar Cycles 23 and 24. In general, geomagnetic storms occur more frequently during the descending phase of these cycles, with a secondary enhancement of this peak at the late ascending phase of the cycle (Gonzalez et al., 1990). Using sunspot numbers might provide a calibration to the models which are trained in particular periods of the solar cycle to provide accurate predictions on other parts of the solar cycle.

The monthly sunspot numbers (smoothed) for the years January 1998 through December 2021, obtained from SWPC data are described in Table 3.

4.4. *Dst* Data

The WDC for Geomagnetism in Kyoto is the official producer of the *Dst* values. We use their *Dst* data at 1-hr intervals from 1998 to 2020 for the challenge. The *Dst* data for the years 1998 through 2014 were "final," 2015 through 2016 were "provisional" and 2017 through 2020 were "real-time." Note that while quicklook data are, at times, unusable in real-time, the WDC corrects them retroactively. Hourly *Dst* values are defined as the true average of the measured values spanning the minute values of 1 hr (Jankowski & Sucksdorff, 1996). Note that the minute values are centered on the whole minute, while the hourly values are centered on the middle. Thus, the first *Dst* value of a day (00:00:00 UTC) covers the measurements between 00:00:30 UTC to 01:00:30 UTC and so forth. When a model is asked to forecast *Dst* values at a specific hour, say at 10:00:00 UTC, the model uses the solar wind data collected up until this time and forecasts the *Dst* at 10:00:00 a.m., which represents the average of ground measurements between 10:00:30 UTC and 11:00:30 UTC. Thus the model forecast is a true forecast, forward in time.

An important step in ML model development and a basic tenet of explainable AI (XAI) is feature exploration. Figure 3 depicts the instantaneous correlation values between sunspot number and solar wind parameters and *Dst* across the complete training data set (Table 4, Figure 4). It is immediately apparent that as expected, |*Dst*| is well correlated with solar wind speed. The highest four correlations with |*Dst*| in this data set are solar wind speed (0.46), IMF magnitude Bt (0.31), temperature (0.25), and IMF Z_{gsm} (0.20). It's also clear that generally, *Dst* is uncorrelated with the locations of the ACE and DSCOVR spacecraft (i.e., "gse_*_ace," "gse_*_dscovr"). As we explore potential ML model architectures and parameterizations, we should expect the most performant models to be generally more sensitive to these parameters and less sensitive to others. We will touch more on post-model ad-hoc XAI in Section 7.

Table 3
Smoothed Sunspot Numbers for January 1998 Through December 2021

Variable ID	Variable	Description	Unit	Min/Max
23	time_delta	Time delta from the start of a segment, for example, 15 days 00:00:00	–	
24	smoothed_ssn	Monthly sunspot numbers, smoothed[a]	Numbers	1.8/180.3

[a]Note that data after April 2020 are predicted by a model.

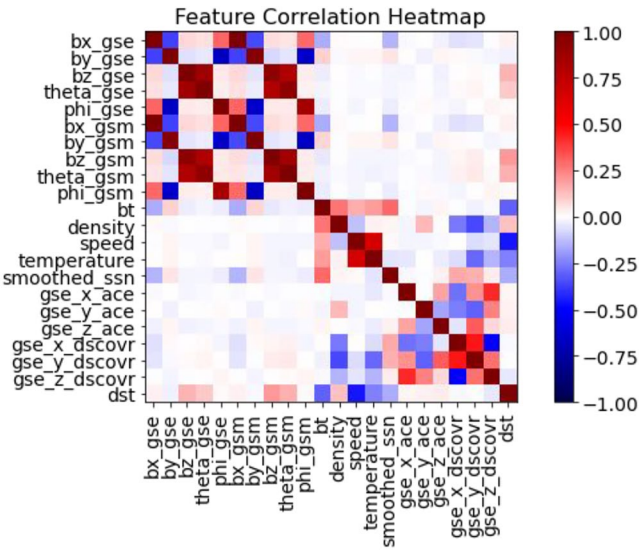

Figure 3. Correlations between sunspot number and solar wind parameters, and *Dst* is shown as a heat map. The color scale shown range is [−1, +1] representing maximally anti-correlated to maximally correlated. The *x*- and *y*-axes are the same, with solar parameters occupying the first 21 columns/rows and *Dst* the last.

5. Data Preparation for the Competition

We divided the data sets into six parts, with only training segments being open to the competitors. The start and end times of each of the subsegments are given in the following Table 4. While dividing the data, we made sure to include energetic parts of the solar cycle in both private and public sections. Within each segment, the original timestamps were replaced with timestamps relative to the beginning timestamps for that section to obfuscate the actual timestamps. Note that this process did not change the order and hence did not violate the causality of the time series.

The competitors used the training part ("train_a," "train_b" and "train_c") data to develop and improve their models. When they submitted a model, the competition platform used the test data sets ("test_a," "test_b," and "test_c") to calculate the accuracy of the model. Figure 4 depicts the sequences of training and test data periods alongside the *Dst* and sunspot time series. To impede competitors from accessing the publicly available *Dst* values, we obfuscated the timestamps. We also took measures to prevent the leakage of chronological information by randomly assigning period names. These precautions were deemed unnecessary for the private test data, as it was only accessible within the code execution harness. The model evaluation was done separately for a public leaderboard and a private leaderboard. The public leaderboard was openly accessible whereas the private leaderboard was restricted to the competition administrators. The data from all of the test sets (a, b, and c) were used on the public leaderboard and private leaderboards. We randomly sampled rows to be included in the public and private leaderboards. Based on relative performance from the public leaderboard as a clue, the teams iterated their models. The private leaderboard provides an unbiased and accurate ranking of the participants' models based on their performance on the withheld test data. Using all data may cause participants to overfit their models to the specific patterns in the public test data. Hence, the final ranking of the models was done on the private leaderboard.

6. Competition

6.1. Benchmark Model

We created a benchmark *Dst* model and a tutorial to aid participants in the competition. The tutorial covers topics such as data correlation, feature selection, and managing gaps and noise in the data. The model uses a simple Long-Short Term Memory (LSTM) neural network architecture, which is explained in Hochreiter and Schmidhuber's (1997) work. You can access the tutorial through this link: https://drivendata.co/blog/model-geomagnetic-field-benchmark.

Table 4
The Beginning and End Timestamps of the Data Sections

Period	Beginning	End
train_a	1998, 2, 16, '00:00:00'	2001, 5, 31, '23:59:00'
train_b	2013, 6, 1, '00:00:00'	2019, 5, 31, '23:59:00'
train_c	2004, 5, 1, '00:00:00'	2010, 12, 31, '23:59:00'
test_a	2001, 6, 1, '00:00:00'	2004, 4, 30, '23:59:00'
test_b	2011, 1, 1, '00:00:00'	2013, 5, 31, '23:59:00'
test_c	2019, 6, 1, '00:00:00'	2020, 10, 31, '23:59:00'

Note. The format for the timestamp is YYYY, MM, DD 'HH:MM:SS'.

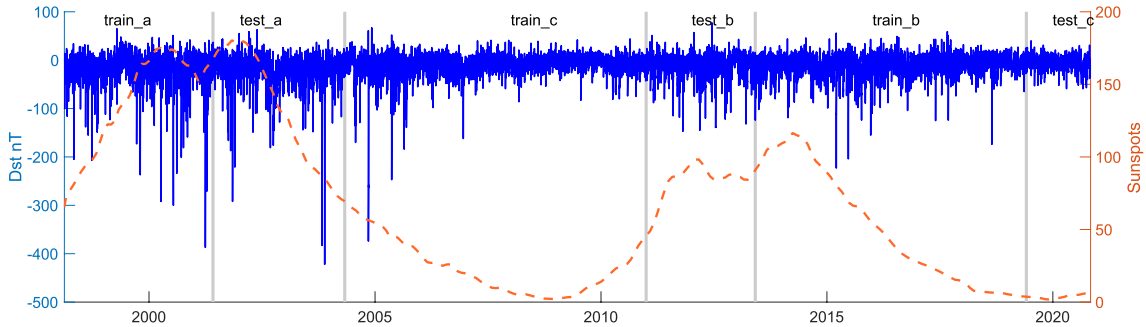

Figure 4. The plot shows the solar activity as the sunspot number (SSN) (red), the geomagnetic storm index *Dst* (blue), and the public and private data segments (top labels). The time range shown is January 1998 through December 2022.

6.2. Real-Time Containerized Testing Environment

For the competition, solvers were required to submit the model files along with the code to make predictions, to a containerized code execution environment. The container had access to four vCPUs and 14 GB RAM but no Graphical Processing Units or network access. The container execution did not have root access to the filesystem. The container was a shared resource, so the solvers were required to be conscientious in their use of resources by adding progress information to their logs and canceling jobs that would run longer than the time limit. All necessary files (forward prediction, model files, etc.) were required to be in the submission. The submissions were to be zip archives named with the extension.zip, containing a "predict.py" file that implements a function "predict_dst" in Python language. This function should be able to make predictions for the current hour t_0 and the hour after that t_{+1} using up to 7 days' worth of RTSW data. While it is acknowledged that the magnetosphere prior to 7 days can have an impact on the predictions, the immediate and recent changes are likely to have a more direct influence on the Dst values. By providing a reasonable window of past data, participants can develop models that offer accurate predictions while still being computationally efficient enough for real-time applications. The solver had to choose a sensible way to handle missing values and noisy data. The models were automatically evaluated upon submission and the scores for public and private leaderboards were generated. The submissions were required to complete the execution within 8 hr, and no single prediction could take more than 30 s.

6.3. Performance Metric

We needed a simple and intuitive error metric that can be used to evaluate the performance of the models in the leaderboard. We chose the root-mean-square error (RMSE) of the residuals between predicted and observed *Dst* values as the metrics for model performance.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{1}$$

where

- $\hat{y}_i$ is the estimated *Dst* values t_0 and t_{+1}
- y_i is the observed *Dst* for t_0 and t_{+1}
- n is the number of samples

While we acknowledge that there exist several error metrics to choose from (multiple skill scores evaluation of *Dst* models is discussed in Rastätter et al., 2013), we have opted for RMSE for several reasons. First, we want to ensure that our model is sensitive to rare, large events, and squaring the errors during RMSE calculation amplifies the impact of such events. This is crucial as large errors can significantly affect the model's accuracy. Additionally, RMSE is widely used in scientific literature, enabling us to compare our model's performance to those previously published. Moreover, RMSE measures the average error magnitude in the dependent variable's units, making it straightforward to interpret model performance. Finally, being a differentiable function, RMSE can be utilized in ML algorithms that depend on gradient descent to optimize their parameters. This makes it a favored loss function for ML algorithms that minimize error through gradient descent.

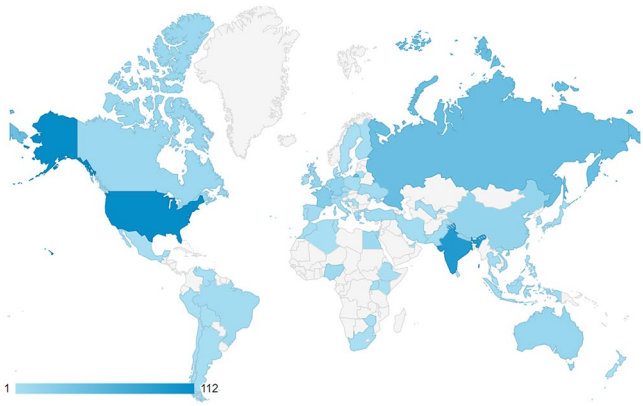

Figure 5. Country affiliations of the 622 participants of the competition.

6.4. Competition Rules

To ensure fairness and adherence to competition rules, DrivenData established the following guidelines for participants. First, signing up for multiple accounts and submitting entries from different accounts was strictly prohibited. Second, privately sharing code or data outside of teams was also not allowed. In addition, winning solutions had to be made available under the MIT License, which is an open-source software license commonly described at https://opensource.org/licenses/MIT, to be eligible for recognition and prize money if offered. Moreover, the use of external data was strictly prohibited, and any attempt to circumvent submission limits resulted in disqualification. Additionally, employees and contractors affiliated with DrivenData and NOAA/CIRES were not eligible to win a prize. Participants were limited to a maximum of three submissions per week. This was partly due to resource constraints in the containerized execution environment, and more significantly to discourage overfitting, encourage local testing, and give participants ample time to iterate and enhance their models. Lastly, the winning solutions had to be documented using a Winning Model Documentation Template provided to top-ranking participants to be eligible for recognition and prize money if offered.

6.5. Competition Progress

The extensive engagement with the data science and space-physics community resulted in 5,448 visitors to the competition site from 105 countries. Out of these visitors, 622 participants joined the challenge from 64 countries (Figure 5). Top countries by the number of participants: USA (22%) India (18%) Russia (9%), France (5%), Germany (3%), China (3%), UK (3%), Australia (3%), Nepal (2%), Canada (2%), other (30%). Since the containerized code submission and execution required higher than usual requirements for results submission, only 112 participants made it all the way through successful code submissions. Each participant (or team) was allowed to submit three models per week, and a total of 1,197 model submissions were generated.

Throughout the competition, the solvers submitted the models to DrivenData's model containers. On successful submissions of a model, two scores are generated. The model is evaluated against a portion of the test data (See Figure 4) to create a score (RMSE error) for the public board and the model is separately evaluated against another part of the test data to create a private scoreboard. The competition winners were selected on the private score. The public scoreboard is visible to all participants, whereas the private scoreboard is only visible to the competition administrators. This arrangement is to encourage the solvers to generalize their model. For example, if the solver is primarily focused on reducing the public score, their model might not perform equally well on the private data set. Figure 6 shows the evolution of the best public and private scores (weekly) over the course of the competition. The scores show a sudden decrease in the second week, followed by a gradual reduction to RMSE 11.1 nT on the private score by the winning model at the end of the competition. One often sees the top score to date improve quickly at the beginning of the competition once participants have had time to implement and train initial modeling approaches, then improve more gradually as they explore the performance boundary given the

Figure 6. Progress of the best scores (root mean square error, the lower the better) on public and private leaderboards over the course of the competition. Benchmark model achieved an root mean square error of 15.2 nT on the private leaderboard and 16.3 on the public leaderboard.

	User or team		Best private RMSE	Timestamp	Trend (last 10)	# Entries
	Ammarali32	1	11.1311	2021-02-03 14:56:37		24
	belinda_trotta	2	11.2532	2021-02-08 08:56:32		20
	LosExtraterrestres	3	11.2944	2021-02-08 04:30:18		17
	k_squared	4	11.5293	2021-02-11 09:56:59		20
	leigh.plt	5	11.6183	2021-02-04 06:26:26		15
	LastRocky	6	11.6504	2021-02-08 07:43:21		6

Figure 7. Final private leaderboard.

data and signal. Another example from a competition to classify penguins is available here https://drivendata.co/blog/aleatoric-limit1.

A snapshot of the final leaderboard is provided in Figure 7. The leaderboard shows the username, private RMSE score, timestamps of their best model submission, and a trend of their RMSE errors over the course of their competition participation. The top winners submitted more than 17 models, achieving final RMSEs of 11.13, 11.25, 11.29, and 11.53 nT in their order of ranking. The very low spread of their scores is typical of data-science competitions, where the ranking is typically determined by fractions of the score values.

7. Winning Models

In the remainder of this paper, we will focus on the top four winning models, selected by their RMSE errors on the private leaderboard. A summary of the models, number of parameters, and model architecture is provided in Table 5.

While all the top models achieved RMSE on the private data between 11.13 and 11.53 nT, they used a diverse set of strategies to achieve that. The models differed in the architecture, input data used, length of the data streams, and the number of parameters they needed to obtain the final results.

7.1. First Place Model Architecture and Data Pre-Processing

The model consists of a Bidirectional LSTM layer followed by a bidirectional Gated Recurrent Unit (GRU, Cho et al., 2014) layer. These layers process each timestep in sequence, and each output is combined with the next step's input. The LSTM and GRU cells contain "forget gates" which allow old data to be discarded when it is no longer relevant while preserving relevant data indefinitely. This allows it to process long sequences without the error gradients "exploding" (becoming very large, due to repeated application of a nonlinear function to an input), which was a problem in older recurrent neural network architectures.

The LSTM architecture is more complex and requires more parameters than the GRU, so a smaller number of units are used in the LSTM to keep the computation feasible.

Table 5
Winning Model Rank and Machine Learning Architecture

Rank	Private score RMSE (nT)	Number of parameters	ML model/method	Input variables used (IDs from Tables 1–3)
1	11.13	60 million	Bidirectional LSTM-GRU; three Flattening Layer; three Dense Layers	2–12,14–16,24
2	11.25	51,191	Ensemble of five convolutional models (CNN)	7,8,9,12,14,15,24
3	11.3	34,354	Ensemble: 1 Light Gradient-boosted Model (LGBM), 2 Feed-forward Neural Networks	7–10,14–16,19–21,24
4	11.53	2.6 million	Ensemble of 21, 4-block deep Convolutional Neural Networks (CNN)	7–10,14–16,19–21,24

Data is aggregated into 1-hr periods, and the mean and standard deviation of the input features are taken for each period. The model uses the 128 hr before prediction time (i.e., around 5 days of data). Missing values are filled using the most frequent value of each feature, and the data is normalized by subtracting the mean and dividing by the standard deviation.

Following the recurrent layers are several dense layers with linear output. Although mathematically these do not change the expressiveness of the model (because any composition of linear functions can be replaced with a single linear function), they may affect the convergence behavior or the initial state.

7.2. Second Place Model Architecture and Data Pre-Processing

The model is a convolutional neural network with an architecture designed to give more importance to later points of the time series, while also capturing larger-scale patterns over the whole series. The network consists of a set of convolutional layers which detect patterns at progressively longer periods. Following all the convolutional layers is a layer that concatenates the last data point of each of the convolution outputs. This concatenation is then fed into a dense layer. The idea of taking the last data point of each convolution is that it represents the patterns at different time spans leading up to the prediction time: for example, the last data point of the first layer gives the features of the hour before the prediction time, then the second layer gives the last 6 hr, etc.

The architecture is somewhat similar to a widely used architecture for image segmentation, the U-Net introduced by Ronneberger et al. (2015). The U-Net consists of a "contracting path," a series of convolutional layers that condense the image, followed by an "expansive path" of up-convolution layers that expand the outputs back to the scale of the original image. Combining small-scale and large-scale features allows the network to make localized predictions that also take account of larger surrounding patterns. The idea is also similar to the Temporal Convolutional Network described by Bai et al. (2018); however, their architecture uses residual (i.e., additive) connections to blend the low-level and high-level features, rather than concatenations.

Missing data is filled by linear interpolation (to reduce noise, the interpolation uses a smoothed rolling average, rather than just the two points immediately before and after the missing part). Features are normalized by subtracting the median and dividing by the interquartile range (this approach is used rather than the more usual mean and standard deviation because some variables have asymmetric distributions with long tails). Data is aggregated in 10-min increments, taking the mean and standard deviation of each feature in the increment.

The final model is an ensemble of five models with the same structure, trained on different subsets of the data. Separate models are trained for times t and $t + 1$, yielding 10 models in total. This ensemble averaging is a common technique in ML. The idea is that each model only imperfectly captures the "true" relationship between the input and output variables, and partly fits noise in the training data. But if we average several models, the random noise components will approximately cancel each other out, leaving a more accurate prediction of the true relationship. We will touch on this idea when we evaluate the models later in the paper.

7.3. Third Place Model Architecture and Data Pre-Processing

The model is an ensemble of a tree-boosting model and two neural networks. There are many engineered features, including some derived from the Fourier transform of the time series. For all three models, the model is trained twice. After the first training, insignificant features were identified and removed before training the final model. Each feature's importance is evaluated by using the model to predict a synthetic data set where that feature's values have been randomly permuted, then measuring the difference in the loss function (a technique used in XAI).

The tree-boosting model is implemented in a Light Gradient Boosting Machine (LightGBM, Ke et al., 2017), while the neural networks are implemented in PyTorch. Both neural networks are dense feed-forward architectures with two layers, using rectified linear activations. The final result is scaled using a sigmoid function so that the minimum and maximum predictions exceed the training data's minimum and maximum by a factor of at most 1.2. The two neural network models are identical except for the number of neurons used in the dense layers (50 and 100).

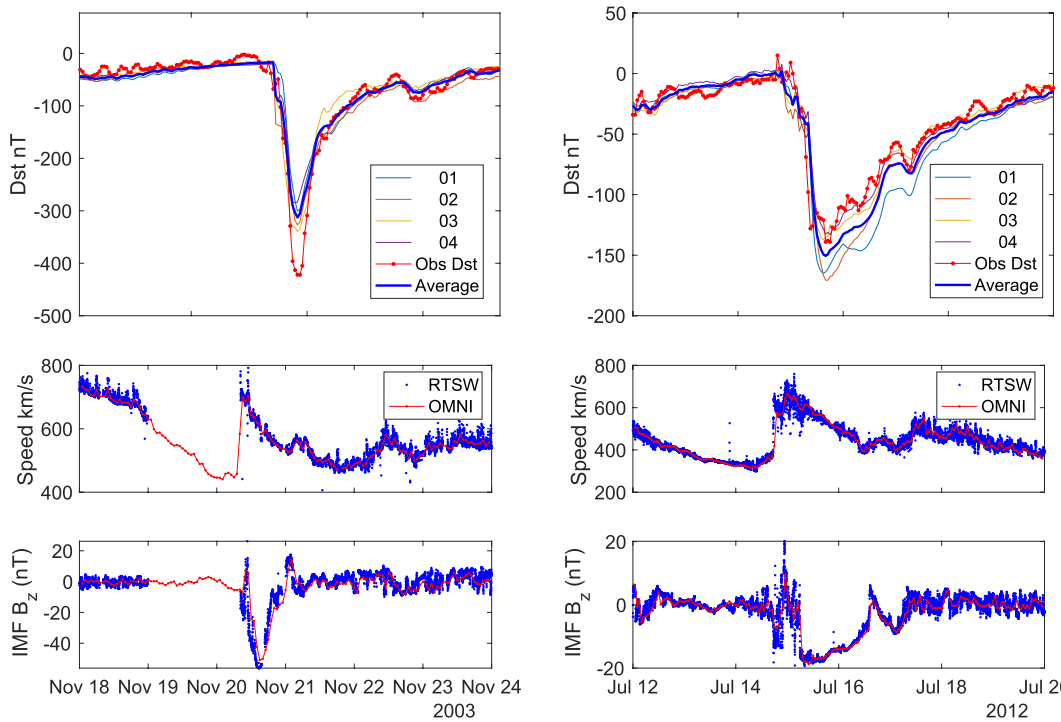

Figure 8. Left panels. A comparison of the observed *Dst* values and model predictions for the 20–23 November 2003 geomagnetic storm. The middle and lower panels show a comparison of Real-Time Solar Wind (RTSW) (1 min) and OMNI (1 hr) data for the same period. The models use the RTSW data. The right panel shows the same data for the 15 July 2012, geomagnetic storm.

7.4. Fourth Place Model Architecture and Data Pre-Processing

The model is an ensemble of 21 convolutional neural network models. The cells have leaky rectified linear activation, and max pooling is used after each convolution to reduce the size of the output. The network has a skip-connection that concatenates the last timestep of the input with the output of the convolutional part of the network, before the final output layers (using a similar idea to the structure used in the second-place model).

Features are aggregated by hour, and the mean and standard deviation are calculated for each period. The last 96 hr (i.e., 4 days) are used in the model. Missing features are filled by interpolation, and data is normalized by subtracting the mean and dividing by the standard deviation.

The models are trained using a custom loss function with parameter p, calculated as follows:

$$\text{Loss} = |y_{\text{predicted}} - y_{\text{true}}| + \left(\log_2\left(\left(y_{\text{predicted}} - y_{\text{true}}\right)^2 + 1\right)\right)^p \quad (2)$$

When $p = 2$, the second term of the loss is similar to the mean squared error. Higher values of p penalize outliers more heavily. This effect arises from the specific formulation of the second term in the loss function. Intuitively, this behavior can be understood as an increased sensitivity to extreme errors. With higher values of p, the loss function assigns more significance to the squared difference between the predicted and true values, amplifying the penalty for outliers. This emphasizes the model's focus on reducing large errors and improving its performance on extreme data points.

The ensemble consists of 21 models trained using the above loss function with p values of 1.5, 2.4, and 2.5; and 7 different random seeds.

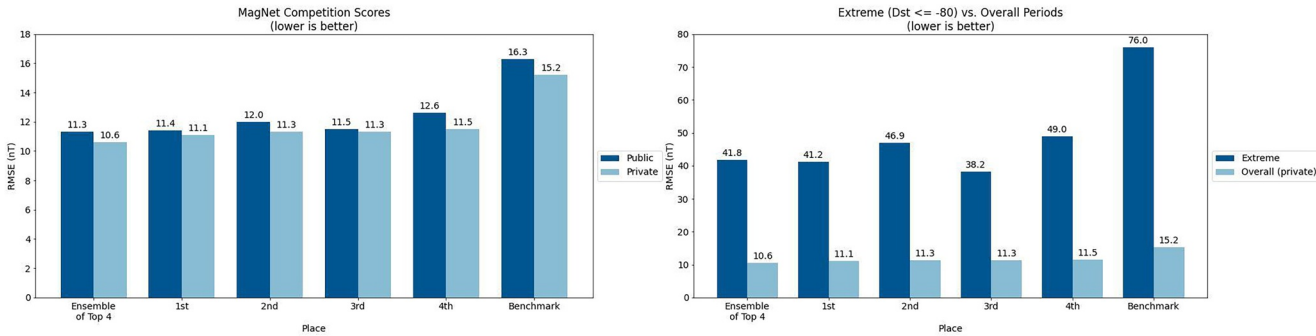

Figure 9. MagNet top models root mean square error (RMSE) scores (vertical axes; lower is better). The left figure shows RMSE on public (dark blue) and private (light blue) for the ensemble of all four top models, each model independently and the original benchmark model. The right figure is the same layout with each model's RMSE score for extreme $Dst < -80$ nT (dark blue) and for overall private test (light blue) periods.

7.5. Performance of the Models

Figure 8, left panel, compares the final *Dst* values during the major geomagnetic storm of 20–23 November 2003, with predictions made by the top four models and their ensemble average. This storm was the largest of solar cycle 23, with a *Dst* index peak value of −422 nT. However, during this event, the Solar Wind Electron Proton Alpha Monitor (SWEPAM; McComas et al., 1998) of the ACE satellite was overwhelmed by high-energy solar particles, and a communication error caused a data gap before the main event (Fernandez-Gomez et al., 2019). Due to a combination of these issues, the RTSW plasma data (density, velocity, and electron temperature) had major gaps lasting more than a day before the onset of the storm and smaller gaps during the evolution of this storm (Posner et al., 2014). The missing plasma data were later reconstructed with the NASA-OMNI post-processing (Fernandez-Gomez et al., 2019; Skoug et al., 2004), introducing clear differences between the operational (RTSW) and research (OMNI) magnetospheric inputs. At the time of the event, the only available data would have been the RTSW data. The models used the RTSW data to predict the *Dst* values, gaps filled by pre-processing or interpolation as defined in Sections 7.2–7.6. Despite these issues, the models performed reasonably well. On the right panels of the figure, model predictions, and key solar wind data are shown for the moderate geomagnetic storm of 14–17 July 2012, during which ACE data was complete throughout the storm period.

Figure 9 presents the overall metrics of the models. The left panel displays the RMSE of the model predictions against the test data in both the public and private leaderboards. The models were ranked according to their performance against the private data. It is worth noting that for the same models, the public data set had slightly larger errors compared to the private data set, which can be attributed to the presence of a greater number of geomagnetic storms in the public data set. It is important to recall that the private and public data sets are distinct from each other, and only the public data set was provided to the competition participants. The average of the top four models had lower errors than any of the individual models, while the benchmark model had significantly higher errors, as expected since it was designed as a beginner model to guide participants. The right panels show the corresponding model performances during geomagnetically active (defined as *Dst* values less than −80 nT) and quiet periods in the private data set. The performance of the models during active periods is slightly different from their overall performance. For instance, the third-place model has the lowest errors (38.2 nT), slightly lower than the models in the first and second places.

7.6. Independent Validation of the Models

The top four models in the challenge were validated by the NCEI/CIRES members before announcing the winners. This step served two main purposes: (a) to ensure that the winning models followed the competition rules and (b) to assess their suitability for NCEI operational use. The validation involved the following components. First, we ensured that the model software and documentation were sufficient for the task. Additionally, we checked the model software for any suspicious activity (e.g., sourcing data external to the data officially provided to all the participants). We then evaluated the models against data collected after the competition (from 20 November 2020 to 18 March 2021) using the RTSW data from SWPC and the Quicklook *Dst* from

Table 6
Winning Model Validation Rubric and Results[a]

Model	Model software and documentation	Check software for suspicious activity	Infer winning models against new data (November '20 to March '21) (RMSE nT)	Train models on public data to reproduce competition RMSE nT (NCEI/Solver provided)	Train models on all competition data (public + private); infer on fresh data	Get training times and inference times (HH:MM NCEI/Solver provided)
First	✓	✓	✓ (5.85)	✓ (11.13/11.13)	✓ (6.07)	✓ (02:56/1:49)
Second	✓	✓	✓ (5.96)	✓ (11.31/11.25)	✓ (5.96)	✓ (00:19/00:16)
Third	✓	✓	✓ (6.43)	✓ (11.35/11.29)	✓ (6.65)	✓ (04:56/02:00)
Fourth	✓	✓	✓ (6.43)	✓ (11.71/11.53)	✓ (6.5)	(12:00/00:40)

[a]Note that the root mean square error (RMSE) numbers refer to the *Dst* prediction at t_0. The marker ✓ indicates a satisfactory outcome.

WDC Kyoto. We found that the relative performance of the models was similar to that on the private leaderboard. We then trained the models on the public data set and reproduced the error metrics, obtaining RMSE errors close to the leaderboard numbers. The small differences could be explained by the random initialization of the model coefficients and the randomness of the optimization method used in the models. Considering these variabilities, it is reassuring to see that the retrained models are producing the same error metrics as in their competition leaderboard. We then trained the models against all the provided data to the competition (private + public) and then evaluated the models against the post-competition data. In this case, we find some changes to the performance ranking with the RMSE errors of the first and third models increasing by 0.22 nT. We also notice that the time taken to train the models by NCEI/CIRES using public data varied significantly from what is documented by the model developers. For the first and third models, the training time doubled. The fourth-place model took 12 hr to complete the training (given a prior estimate of 40 min). However, the inference time for a single-step prediction (t_0 and t_{+1}) took only less than a minute for all the models. Table 6 summarizes our validation results.

7.7. Post-Competition Evaluation of the Models

A more meaningful evaluation of the model is to evaluate them over a sufficiently long time series collected after the competition phase. Ideally, this period of evaluation should have several large geomagnetic storms resulting in several periods where *Dst* values are less than −100 nT. However, as we write this paper, we are at the beginning of the solar cycle 25 with rare occurrences of larger geomagnetic storms. We use RTSW data and *Dst* ("Quicklook") collected after the conclusion of the competition for a second evaluation of the top four models. Specifically, we use the data collected from March 2021 through May 2022. The models were used to predict 9,990 hourly values of *Dst* using the RTSW data. Provisional *Dst* data were available up to the end of 2021, and for the rest of the period, we used the Quicklook *Dst* data. In this period, we had five geomagnetic storms with *Dst* values below −80 nT and the biggest storm was on 4 November 2021. As an external benchmark model, we chose the empirical model "LASP" by Temerin and Li (2002) for the following reasons. The "LASP" model does not use historical *Dst* values to predict future *Dst* values. It is solely reliant on solar wind data for *Dst* prediction. We downloaded the real-time model values from https://lasp.colorado.edu/space_weather/dsttemerin/dsttemerin.html and they were interpolated to the hourly timestamps of the observed *Dst* values.

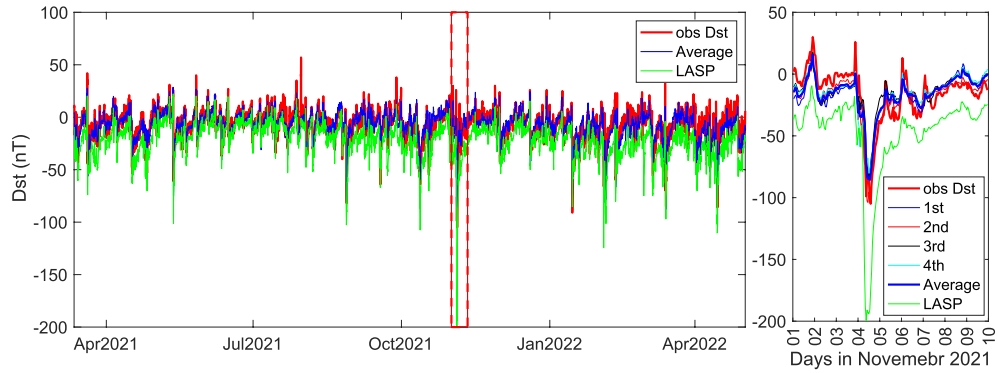

Figure 10. Left panel shows *Dst* observed (blue-green) and predicted by models (blue—an average of four models, green-LSAP) over the post-competition period from March 2021 through May 2022. The right panel is zoomed (area in the red rectangle) on November 1-10, 2021.

Figure 10 shows the model predictions and observations for the period of the evaluation. The models generally predict the observed *Dst* variations correctly. The largest geomagnetic storm happened on 5 November 2021. The MagNet models follow the observed *Dst* values closely in this period. The LASP model shows a large negative bias (average mean error of −15.5 nT), as well as over prediction of the peak *Dst* on 4 November 2021, by about 100 nT. A part of the bias error of the LASP model may be explained by the fact that we are comparing the "QuickLook" *Dst*, which might be adjusted when provisional or final versions are released.

The data and predictions by the top four models are given in Table 7. For presenting the metrics, we chose to separate the RMSE error into standard deviation and mean since the quiet-time baselines of Quicklook *Dst* values are not considered final and we need to determine the mean errors of the models separately. The models generally predict the *Dst* variations at t_0 with standard deviations in the range of 7.02–7.64 nT, with very small mean errors. The models have slightly larger errors at $t1$ with a standard deviation range of 7.13–7.79 nT. The predictions are highly correlated with the observed *Dst* data. The ensemble average of the four models has the lowest standard deviation error at t_0 and t_{+1} with the observations. The average model also shows the highest correlation coefficients. This result is consistent with previous studies (e.g., Bojer & Meldgaard, 2021) that showed that ensembles of diverse ML models have better prediction accuracy than any individual models. The top four models used three different Machine-Learning approaches and data preprocessing methods (see Table 5 and Sections 7.1–7.4). Furthermore, the second and fourth models themselves are ensemble averages of 5 and 4 convolutional neural networks respectively. We speculate that diverse modeling approaches allow for the ensemble to see more aspects of the phenomenon that we try to model than any individual models. The residuals of the LASP model show significantly larger standard deviation and mean values. Additionally, the model has a smaller correlation with the observed data.

Table 7
Metrics for Post-Challenge Model Evaluation[a]

	Prediction at t_0			Prediction at t_1		
Model	σ (nT)	μ (nT)	ρ (x,y)	σ (nT)	μ (nT)	ρ (x,y)
Average (1–4)	6.77	0.97	0.87	6.94	1.46	0.87
First	7.64	1.49	0.85	7.79	1.28	0.84
Second	7.02	1.75	0.87	7.38	0.69	0.86
Third	7.26	0.67	0.85	7.33	0.23	0.85
Fourth	7.04	−0.023	0.86	7.13	0.91	0.85
LASP	9.46	15.528	0.82	–	–	–

[a]Metrics for post-challenge evaluation of models used coefficients provided by the solvers. "σ" reflects the standard deviation of the residuals in nanotesla (lower is better), and "μ" reflects the mean of the residuals in nanotesla. "ρ" reflects the Pearson correlation coefficient between samples (higher is better).

To better understand the models' post-competition performance, their residuals were binned and their errors were examined across different levels of geomagnetic activity. In Figure 11 the standard deviation errors are shown as a function of magnetic activity, using Dst bin boundaries of (25,0,−5,−10,−15,−20,−25,−50,−75,−125), with the number of hourly data points available in each bin being (2957,1462,1241,1002,639,432,684,83,19) respectively. Generally, the errors of the models increase as the level of geomagnetic activity rises. For predictions at t_0 (left panels) the LASP model had comparable errors with the MagNet models for *Dst* bins centered at −37.5 nT and higher. However, the LASP model (green) had significantly higher errors for more geomagnetically active periods. The first and second models show slightly larger errors than the rest of the MagNet models for moderately active periods (−37.5 to −62.5 nT) but show signs of reduced errors for the most active periods (*Dst* bin centered on −100 nT). We can make similar observations for

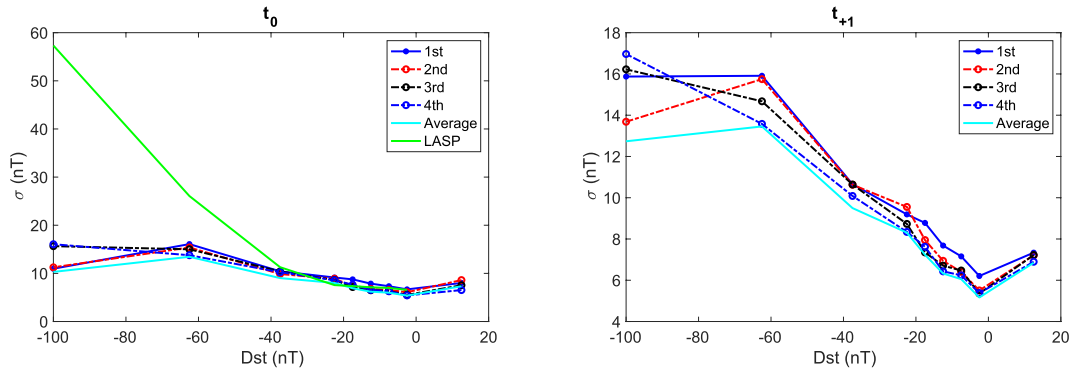

Figure 11. Depicts the root mean square error score (vertical axis) for each of the top four models versus *Dst* (horizontal axis). The traces are first place—solid blue, second—red, third—black, fourth—dashed blue, and ensemble mean—light blue. The left figure is prediction at t_0 and the right figure is at t_{+1}.

predictions at t_{+1}, with slightly larger errors for all periods. The average of the MagNet ensemble (cyan) has the lowest errors in all periods and for both t_0 and t_{+1}.

As briefly introduced in Section 4, XAI is a critical component of ML model development as we strive to understand the physical ramifications of a given trained model, that is, its behavior when presented with data it wasn't trained on, ultimately avoiding issues associated with "black box" ML models (e.g., McGovern et al., 2019). Insufficiently interpretable models are highly unlikely to gain sufficient trust with the target user community to be deployed into an operational or decision-making environment. XAI is an active area of research, with new explainability tools being rapidly developed by community efforts, such as the scikit-explain Python package (e.g., Chase, Harrison, Burke, et al., 2022; Chase, Harrison, Lackmann, & McGovern, 2022; Flora et al., 2022a, 2022b). The independent evaluations discussed in Sections 7.5 and 7.6 are one facet of explainability. Another is a set of global model agnostic interrogations (e.g., Molnar, 2020 their Section 8.5; Fisher et al., 2018). In Figure 12, we have employed a technique referred to as *permutation feature importance*, whereby the inputs (features) are permuted individually, to break the relationship between the input (feature) and the correct *Dst* (outcome). This figure demonstrates the

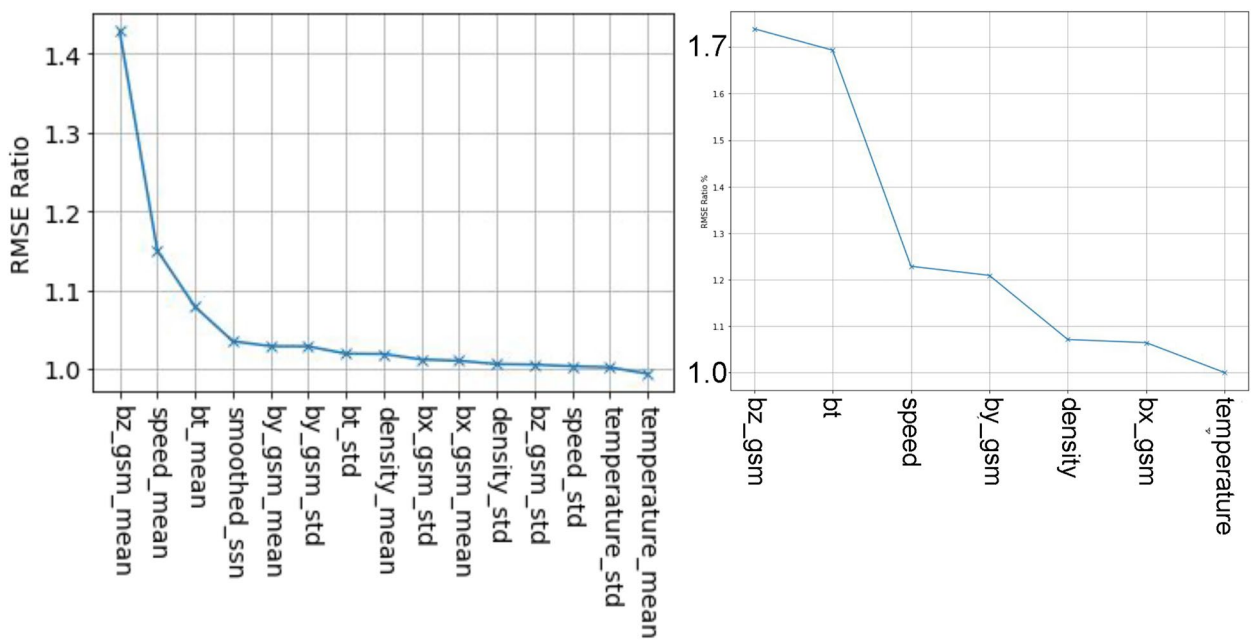

Figure 12. Feature permutation importance plots for the benchmark Long-Short Term Memory model (left) and the second place model (right). The *y*-axis is the ratio of the root mean square error (RMSE) for each feature to the most important feature ("bz_gsm" in both cases).

relative effect different input parameters have on the model's ability to predict *Dst*. We show two cases, the feature importance for the benchmark LSTM model versus the second place model. In both cases, the top three most important inputs are the z-component of the interplanetary magnetic field (IMF Bz), solar wind speed, and the total magnetic field (IMF Bt).

8. Summary and Conclusion

NOAA's NCEI, in collaboration with the University of Colorado's CIRES and NASA's CoECI, organized an open data-science challenge named "MagNet: Model the Geomagnetic Field" to predict *Dst* in a setup that can be used operationally. DrivenData and HeroX implemented the challenge, which drew 622 participants from 64 countries, resulting in 1,197 model submissions using various ML approaches. The models were automatically evaluated against two portions of solar wind and Dst data using a containerized computing system that simulated a real-time modeling environment. The top four winning models, chosen based on their lowest RMSE errors, were further evaluated in this paper. These models used different modeling architectures, and the ensemble averages of their outputs performed better than individual models for both competition and post-competition data. The models were robust against data corruption and outages affecting the ACE plasma sensors during the November 2003 geomagnetic storm. NCEI re-evaluated the MagNet models by retraining them and evaluating their operational performance. The models' performance was comparable to the competition leaderboard metrics. Using RTSW data collected after the competition (March 2021 to May 2022), the models' performance was further evaluated against the Dst data. We find that the models generally predict the *Dst* variations at t_0 with standard deviations in the range of 7.02–7.64 nT, with very small mean errors. The models have slightly larger errors at t_1 with a standard deviation range of 7.13–7.79 nT. Again, the ensemble average scored the least errors against the observations. We hope that the MagNet models will serve as a benchmark for improving Dst value forecasts for near real-time operational needs and encourage discussions of new model architectures and data pre-processing techniques. We have also provided examples of how to create a wrapper around the model so that it can run in a real-time containerized environment.

The competition revealed several lessons learned, with notable successes and areas for improvement. Among the successes were the well-organized and cleanly prepared data, which enabled broad participation aided by challenge coverage, outreach, and useful domain resources provided. Additionally, the benchmark blog post used in several winning models was effective, while submission acceptance, containerized execution, and scoring were smooth, allowing for the evaluation of over 1,000 models. Steps were taken to discourage misuse of the public test data by the obfuscation of time stamps to index, code evaluation, and logs, and testing on unseen data. The winner solution validation and sharing were also streamlined by submission requirements. However, opportunities for improvement were identified, such as continuing to test models on unseen data, augmenting post-challenge testing with data containing more geomagnetic storms, and making benchmark resources more ready to use for participant iteration. Finally, the containerized code execution environment proved challenging for some participants, resulting in fewer successful submissions.

The primary objective of the competition was to predict Dst values, but the winning model has the potential to serve additional purposes for the Space-Physics and Industry communities. By retraining the models using RTSW data and the desired output variable, the model can be utilized to forecast other Space-Weather indices.

Data Availability Statement

The models, data, and associated documentation are available from the following websites.

1. The model software for winning solutions from MagNet challenge are publicly available at Ali et al. (2021).
2. The data set used for the competition are available at Nair (2023).
3. Competition website
   - Main https://www.drivendata.org/competitions/73/noaa-magnetic-forecasting/.
   - Benchmark model https://drivendata.co/blog/model-geomagnetic-field-benchmark/.
4. Tutorials designed based on the winning and benchmark models to help students and early-career researchers to explore and improve the models are available at Belinda Trotta (2023).

Acknowledgments
Our heartfelt appreciation goes out to the 600+ participants of the competition whose enthusiasm and diverse strategies were instrumental in achieving the competition's goals. We would like to extend our congratulations and gratitude to the winners, Ammar Ali (first), Belinda Trotta (second), team "Los Extraterrestres" with members Yanick Medina and Hamlet Medina (third), and team "*k*-squared" with members Kareem Eissa and Karim Amer (fourth), for generously sharing their modeling approach and software. We would also like to thank Christine Jenkins and Steve Rader of NASA's Center of Excellence for Collaborative Innovation (CoECI) for their assistance in selecting the competition provider, as well as Michael Husler of NOAA's Space Weather Prediction Center for providing the RTSW data. HeroX provided outreach support. We are grateful to the NCEI Innovates program for partially funding this program and to Eric Kihn and the NCEI Budget team for facilitating the Inter-Agency Agreement between NOAA and NASA. Finally, we acknowledge the support of the NOAA cooperative agreement NA17OAR4320101 for this research.

References

Ali, A., Trotta, B., Medina, Y., Medina, H., Eissa, K., & Amer, K. (2021). Winning code from the MagNet: Model the Geomagnetic Field challenge (Version 1.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.8329881

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Bala, R., & Reiff, P. (2012). Improvements in short-term forecasting of geomagnetic activity. *Space Weather*, *10*(6), S06001. https://doi.org/10.1029/2012sw000779

Bogucki, R., Cygan, M., Khan, C. B., Klimek, M., Milczek, J. K., & Mucha, M. (2019). Applying deep learning to right whale photo identification. *Conservation Biology*, *33*(3), 676–684. https://doi.org/10.1111/cobi.13226

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, *37*(2), 587–603. https://doi.org/10.1016/j.ijforecast.2020.07.007

Boukabara, S.-A., Krasnopolsky, V., Penny, S. G., Stewart, J. Q., McGovern, A., Hall, D., et al. (2020). Outlook for exploiting artificial intelligence in the earth and environmental sciences. *Bulletin America Meteorology Social*, *102*(5), 1–53. https://doi.org/10.1175/bams-d-20-0031.1

Burton, R. K., McPherron, R. L., & Russell, C. T. (1975). An empirical relationship between interplanetary conditions and Dst. *Journal of Geophysical Research*, *80*(31), 4204–4214. https://doi.org/10.1029/ja080i031p04204

Chandorkar, M., Camporeale, E., & Wing, S. (2017). Probabilistic forecasting of the disturbance storm-time index: An autoregressive Gaussian process approach. *Space Weather*, *15*(8), 1004–1019. https://doi.org/10.1002/2017SW001627

Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., & McGovern, A. (2022). A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Weather and Forecasting*, *37*(8), 1509–1529. https://doi.org/10.1175/waf-d-22-0070.1

Chase, R. J., Harrison, D. R., Lackmann, G., & McGovern, A. (2022). A Machine learning tutorial for operational meteorology, Part II: Neural networks and deep learning. arXiv.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Christine (2023). manojcnair/magnet-geomagnetic-field: MagNet competition winning solutions (v1.0.0) [Software]. Zenodo. https://doi.org/10.5281/zenodo.8197485

Chulliat, A., Vigneron, P., & Hulot, G. (2016). First results from the Swarm dedicated ionospheric field inversion chain. *Earth Planets and Space*, *68*(1), 104. https://doi.org/10.1186/s40623-016-0481-6

Chulliat, A., Vigneron, P., Thébault, E., Sirol, O., & Hulot, G. (2013). Swarm SCARF dedicated ionospheric field inversion chain. *Earth Planets and Space*, *65*(11), 8–1283. https://doi.org/10.5047/eps.2013.08.006

Codrescu, M. V., Negrea, C., Fedrizzi, M., Fuller-Rowell, T. J., Dobin, A., Jakowsky, N., et al. (2012). A real-time run of the Coupled Thermosphere Ionosphere Plasmasphere Electrodynamics (CTIPe) model. *Space Weather*, *10*(2). https://doi.org/10.1029/2011sw000736

Cristoforetti, M., Battiston, R., Gobbi, A., Iuppa, R., & Piersanti, M. (2022). Prominence of the training data preparation in geomagnetic storm prediction using deep neural networks. *Scientific Reports*, *12*(1), 7631. https://doi.org/10.1038/s41598-022-11721-8

Fernandez-Gomez, I., Fedrizzi, M., Codrescu, M. V., Borries, C., Fillion, M., & Fuller-Rowell, T. J. (2019). On the difference between real-time and research simulations with CTIPe. *Advances in Space Research*, *64*(10), 2077–2087. https://doi.org/10.1016/j.asr.2019.02.028

Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Retrieved from http://arxiv.org/abs/1801.01489

Flora, M., Potvin, C., McGovern, A., & Handler, S. (2022a). Comparing explanation methods for traditional machine learning models Part 1: An overview of current methods and quantifying their disagreement. Arxiv. https://doi.org/10.48550/arxiv.2211.08943

Flora, M., Potvin, C., McGovern, A., & Handler, S. (2022b). Comparing explanation methods for traditional machine learning models Part 2: Quantifying model explainability faithfulness and improvements with dimensionality reduction. Arxiv. https://doi.org/10.48550/arxiv.2211.10378

Gonzalez, W. D., Gonzalez, A. C., & Tsurutani, B. T. (1990). Dual-peak solar cycle distribution of intense geomagnetic storms. *Planetary and Space Science*, *38*(2), 181–187. https://doi.org/10.1016/0032-0633(90)90082-2

Graham, B. (2015). *Kaggle diabetic retinopathy detection competition report* (pp. 24–26). University of Warwick.

Gruet, M. A., Chandorkar, M., Sicard, A., & Camporeale, E. (2018). Multiple-hour-ahead forecast of the Dst index using a combination of long short-term memory neural network and Gaussian process. *Space Weather*, *16*(11), 1882–1896. https://doi.org/10.1029/2018SW001898

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jankowski, J., & Sucksdorff, C. (1996). Guide for magnetic measurements and observatory practice.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (pp. 3149–3157). Curran Associates Inc.

Laperre, B., Amaya, J., & Lapenta, G. (2020). Dynamic time warping as a new evaluation for Dst forecast with machine learning. *Frontiers in Astronomy and Space Sciences*, *7*, 39. https://doi.org/10.3389/fspas.2020.00039

Lazzús, J. A., Vega, P., Rojas, P., & Salfate, I. (2017). Forecasting the Dst index using a swarm-optimized neural network. *Space Weather*, *15*(8), 1068–1089. https://doi.org/10.1002/2017SW001608

Lundstedt, H., Gleisner, H., & Wintoft, P. (2002). Operational forecasts of the geomagnetic Dst index. *Geophysical Research Letters*, *29*(24), 34-1–34-4. https://doi.org/10.1029/2002gl016151

Manoj, C., & Maus, S. (2012). A real-time forecast service for the ionospheric equatorial zonal electric field. *Space Weather*, *10*(9), S09002. https://doi.org/10.1029/2012sw000825

Martinez, M. G., & Walton, B. (2014). The wisdom of crowds: The potential of online communities as a tool for data analysis. *Technovation*, *34*(4), 203–214. https://doi.org/10.1016/j.technovation.2014.01.011

Maus, S., & Lühr, H. (2005). Signature of the quiet-time magnetospheric magnetic field and its electromagnetic induction in the rotating Earth. *Geophysical Journal International*, *162*(3), 755–763. https://doi.org/10.1111/j.1365-246X.2005.02691.x

Maus, S., Manoj, C., Rauberg, J., Michaelis, I., & Lühr, H. (2010). NOAA/NGDC candidate models for the 11th generation International Geomagnetic Reference Field and the concurrent release of the 6th generation Pomme magnetic model. *Earth Planets and Space*, *62*(10), 729–735. https://doi.org/10.5047/eps.2010.07.006

Maus, S., & Weidelt, P. (2004). Separating the magnetospheric disturbance magnetic field into external and transient internal contributions using a 1D conductivity model of the Earth. *Geophysical Research Letters*, *31*(12), L12614. https://doi.org/10.1029/2004gl020232

McComas, D. J., Bame, S. J., Barker, P., Feldman, W. C., Phillips, J. L., Riley, P., & Griffee, J. W. (1998). Solar wind electron proton alpha monitor (SWEPAM) for the advanced composition explorer. In *The advanced composition explorer mission* (pp. 563–612).

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin America Meteorology Social*, *100*(11), 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Leanpub.

Mursula, K., Qvick, T., Holappa, L., & Asikainen, T. (2022). Magnetic storms during the space age: Occurrence and relation to varying solar activity. *Journal of Geophysical Research: Space Physics*, *127*(12), e2022JA030830. https://doi.org/10.1029/2022JA030830

Nair, M. (2023). Data used for MagNet competition [Dataset]. https://doi.org/10.5281/zenodo.8197443

Nair, M., Chulliat, A., Woods, A., Alken, P., Meyer, B., Poedjono, B., et al. (2021). Next generation high-definition geomagnetic model for wellbore positioning, incorporating new crustal magnetic data. In *Paper presented at the offshore technology conference, Virtual and Houston, Texas, August*. Paper Number: OTC-31044-MS. https://doi.org/10.4043/31044-MS

NASA. (2022). Space physics data facility OMNIWeb interface. Retrieved from https://omniweb.gsfc.nasa.gov/html/ow_data.html

NOAA Artificial Intelligence Strategic Plan. (2021). Retrieved from https://sciencecouncil.noaa.gov/Portals/0/Artificial Intelligence Strategic Plan_Final Signed.pdf?ver=2021-01-19-114254-380

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

Papitashvili, N. E., & King, J. H. (2020). *OMNI 1-min data*. NASA Space Physics Data Facility. https://doi.org/10.48322/45bb-8792

Posner, A., Hesse, M., & St. Cyr, O. C. (2014). The main pillar: Assessment of space weather observational asset performance supporting nowcasting, forecasting, and research to operations. *Space Weather*, *12*(4), 257–276. https://doi.org/10.1002/2013sw001007

Raeder, J., McPherron, R. L., Frank, L. A., Kokubun, S., Lu, G., Mukai, T., et al. (2001). Global simulation of the geospace environment modeling substorm challenge event. *Journal of Geophysical Research*, *106*(A1), 381–395. https://doi.org/10.1029/2000JA000605

Rastätter, L., Kuznetsova, M. M., Glocer, A., Welling, D., Meng, X., Raeder, J., et al. (2013). Geospace environment modeling 2008–2009 challenge: Dst index. *Space Weather*, *11*(4), 187–205. https://doi.org/10.1002/swe.20036

Riley, P., Linker, J. A., & Mikić, Z. (2013). On the application of ensemble modeling techniques to improve ambient solar wind models. *Journal of Geophysical Research: Space Physics*, *118*(2), 600–607. https://doi.org/10.1002/jgra.50156

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Skoug, R. M., Gosling, J. T., Steinberg, J. T., McComas, D. J., Smith, C. W., Ness, N. F., & Burlaga, L. F. (2004). Extremely high speed solar wind: 29–30 October 2003. *Journal of Geophysical Research*, *109*(A9), A09102. https://doi.org/10.1029/2004ja010494

Sugiura, M. (1963). Hourly values of equatorial Dst for the IGY (No. NASA-TM-X-55238).

Tasistro-Hart, A., Grayver, A., & Kuvshinov, A. (2021). Probabilistic geomagnetic storm forecasting via deep learning. *Journal of Geophysical Research: Space Physics*, *126*(1), e2020JA028228. https://doi.org/10.1029/2020ja028228

Temerin, M., & Li, X. (2002). A new model for the prediction of Dst on the basis of the solar wind. *Journal of Geophysical Research*, *107*(A12), SMP31-1–SMP31-8. https://doi.org/10.1029/2001ja007532

Tóth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., De Zeeuw, D. L., et al. (2005). Space weather modeling framework: A new tool for the space science community. *Journal of Geophysical Research*, *110*(A12), A12226. https://doi.org/10.1029/2005JA011126

Trotta, B. (2023). CIRES-Geomagnetism/MagNet: Second place winning solution for MagNet competition (v1.1) [Software]. Zenodo. https://doi.org/10.5281/zenodo.8157282

Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002502. https://doi.org/10.1029/2021MS002502

Zwickl, R. D., Doggett, K. A., Sahm, S., Barrett, W. P., Grubb, R. N., Detman, T. R., et al. (1998). The NOAA real-time solar-wind (RTSW) system using ACE data. In *The advanced composition explorer mission* (pp. 633–648).